110p. mfs N64-17194-N64-17202 NsG-487 *

CODE-1

CR-53170

PROCEEDINGS . . .

SPACE COMMUNICATIONS INSTITUTE

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
U. S. A.

9

JUNE 23-28, 1963

OTS PRICE

XEROX $ 9.10 ph

THIS IS A UNIVERSITY COLLEGE PUBLICATION

UNIVERSITY OF MARYLAND ■ COLLEGE PARK, MARYLAND

U.S.A.

t: PROCEEDINGS OF THE

SPACE COMMUNICATIONS INSTITUTE ,

conducted at the
University of Maryland
College Park, Maryland

by the
Conferences and Institutes Division, University College
and the
Physics and Astronomy Department, College of Arts and Sciences

in cooperation with the
National Aeronautics and Space Administration (NASA)
Washington, D. C.

June 23 - 28, 1963

Richard H. Stottler
Assistant Dean and Director
Conferences and Institutes Division
University College

Frederick J. Tischer
Director
Space Communications Institute

John S. Toll
Professor and Chairman
Physics and Astronomy Department
College of Arts and Sciences

Clive C. Veri
Conference Coordinator
Conferences and Institutes Division
University College

PREFACE

This volume presents the proceedings of the Space Communications Institute, and includes the digests and abstracts of technical papers which were presented at the Institute.

In an effort to offer advanced technical training in space science and technology, the University of Maryland, through the University College and College of Arts and Sciences, joined with the National Aeronautics and Space Administration in conducting the Space Communications Institute. The program afforded participants an opportunity to meet and hear outstanding authorities in various areas of space communications. These distinguished scientists presented the fundamentals, and the recent developments and modern theoretical methods in the field.

The Institute was designed to give scientists and engineers working in one particular specialized field an opportunity to gain a limited theoretical working knowledge in neighboring fields and to inform them of new principles in related fields. Persons involved in the solution of complex space communications problems were given an opportunity to acquire selective basic knowledge condensed from large subject areas. The Institute was further designed to provide a fertile ground for the lively interchange of ideas between the participants.

ACKNOWLEDGEMENT

University College was pleased to join with the College of Arts and Sciences and the National Aeronautics and Space Administration in offering the 1963 Space Communications Institute. We take this opportunity to acknowledge the individuals who merit special citation for their efforts in helping to make the program a success.

We gratefully acknowledge Mr. James C. Reese, Head of the Employee Development Branch at the Goddard Space Flight Center (NASA) for his assistance in initiating the Institute.

Our deep appreciation is extended to Dr. Frederick J. Tischer, Assistant Director of the Research Institute at the University of Alabama, for assuming the duties as Director of the Institute by special appointment. Dr. Tischer capably headed the Advisory Committee, edited the Proceedings, and in general, provided the overall guidance for the program.

We also wish to express our appreciation to the members of the Advisory Committee who gave direction to the content of the Institute, and to the lecturers who devoted a great deal of effort in implementing the desires of the Committee.

We are especially grateful to the National Aeronautics and Space Administration for their financial subsidy of the Institute through Research Grant NsG-482.

Ray Ehrensberger, Dean
University College
November, 1963

INSTITUTE ADVISORY COMMITTEE

Representing the University of Maryland

Dr. F. J. Tischer	Director Space Communications Institute, and Assistant Director Research Institute University of Alabama
Mr. Clive C. Veri	Institute Coordinator Conferences and Institutes Division University College
Dr. Howard J. Laster	Associate Chairman Physics and Astronomy Department
Dr. S. Fred Singer	Director National Weather Satellite Center U. S. Weather Bureau, and Professor (on leave) Physics and Astonomy Department

Representing the National Aeronautics and Space Administration

Mr. James C. Reese	Head Employee Development Branch Goddard Space Flight Center
Mr. Charles Jones	Employee Development Specialist Goddard Space Flight Center
Mr. Varice Henry	SYNCOM II Ground Station Manager Communications Branch Goddard Space Flight Center
Dr. Richard Lehnert	Senior Staff Member Systems Analysis Office Goddard Space Flight Center
Dr. R. Rochelle	Head Flight Data Systems Branch Goddard Space Flight Center
Mr. Curtis Staut	Senior Engineer Space Data Acquisition Division Goddard Space Flight Center
Dr. Frederick Vonbun	Head Systems Analysis Office Goddard Space Flight Center

CONTENTS Page

INTRODUCTION

A NEW APPROACH

by

F. J. Tischer

"Space Communications," in a wider sense of this term, deals with all operations in space science and engineering where radio is involved. With this definition, radio and radar astronomy, telemetry, communication by radio links, tracking, radio navigation and guidance, and space exploration by radio belong in this category of space work. The theoretical fundamentals of all these subjects areas, as far as radio is concerned, are the same.

Looking at these subject areas from this point of view is somewhat unusual. According to the usual concept, scientists and engineers are educated and trained in the fundamentals of these subject areas first disregarding radio; radio, as a means for carrying out the operations, is introduced later as a special case. According to the concept of the Institute, we consider radio communications as the primary subject, and the applications, such as astronomy, tracking, navigation, etc., as special cases in a secondary phase of the consideration. According to this concept, the considerations attain the form of a multidisciplinary activity.

Based on this concept, the Institute brought together scientists and educators specialized in the above subject areas to present and to discuss the common fundamentals and topics of their specialization. The material of the lectures, modern fundamentals as well as specialized topics, was carefully selected from the viewpoint of its usefulness for space work.

The fundamentals common for the subject areas of Space Communications were presented in lectures on electromagnetics, wave propagation, antennas, and communication and information theory.

The specialized topics were presented in groups. Radio astronomy, wave propagation in and properties of the atmosphere, ionosphere, and particle belts were the topics of one of these. Sophisticated coding schemes and decision detection formed another group of lectures. Communication by light beams, made practical by the use of lasers, and related phenomena, specific cases of wave generation, transmission, and detection, tracking, and a comparison of satellite communication systems were presented in other groups as indicated in the table of contents.

The basic concept of the Institute, if strictly carried through, leads to a new type of specialization which, as a complement to the common type, has become necessary in dealing with complex space problems.

The new type of specialist of Space Communications has a thorough knowledge and working ability in the radio- and communication-theoretical areas with a broad knowledge in their application to space work in the various subject areas listed at the beginning. These specialists complement scientists and engineers which have a knowledge and working ability narrowly specialized with regard to specific applications such as astronomy, navigation, communication, etc. Scientists and engineers with the new type of specialization are highly in demand for work dealing with the overall operation of communication systems, for the related analysis work, and for optimization studies.

The multidisciplinary approach followed in setting up such an Institute requires new methods which only by experience and by exploration of the reaction of the audience can be gained. Also from this viewpoint, the Institute was extremely valuable.

FUNDAMENTAL LECTURES

ELECTROMAGNETICS AND ANTENNAS

by Roger F. Harrington

I. Waves

Let us begin with Maxwell's equations as they apply in vacuum:

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad (1) \qquad\qquad \nabla \cdot \vec{E} = 0 \qquad (3)$$

$$\nabla \times \vec{B} = \frac{1}{c^2}\frac{\partial \vec{E}}{\partial t} \qquad (2) \qquad\qquad \nabla \cdot \vec{B} = 0 \qquad (4)$$

where c is the velocity of light. The field quantities $\vec{E}$ (electric intensity) and $\vec{B}$ (magnetic induction) are measurable by a test charge q according to the Lorentz force law

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) \qquad (5)$$

where $\vec{v}$ is the velocity of q. If we take the curl of (1) and substitute from (2), we have

$$\nabla \times \nabla \times \vec{E} = -\frac{1}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} \qquad (6)$$

An application of the vector identity $\nabla \times \nabla \times \vec{E} = \nabla(\nabla \cdot \vec{E}) - \nabla^2 \vec{E}$ gives the usual wave equation

$$\nabla^2 \vec{E} = \frac{1}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} \qquad (7)$$

where ∇^2 is the Laplacian operator. Analogous operations on (1) to (4) give

$$\nabla^2 \vec{B} = \frac{1}{c^2}\frac{\partial^2 \vec{B}}{\partial t^2} \qquad (8)$$

Hence, both $\vec{E}$ and $\vec{B}$ satisfy the vector wave equation in vacuum.

To illustrate the simplest type of wave, consider an electric field which has only one component, say an x component, and is a function only of time and of z (perpendicular to x). Equation (7) then reduces to

$$\frac{\partial^2 E_x}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E_x}{\partial t^2} \qquad (9)$$

This is a one-dimensional wave equation which has two independent solutions

$$E_x = f(t - z/c) + g(t + z/c) \qquad (10)$$

where f and g are arbitrary functions. The first term of (10) represents a wave traveling in the +z direction with no change in shape, and the second term represents a wave traveling in the -z direction with no change in shape.

Let us now restrict consideration to sinusoidally time-varying quantities according to the usual convention

$$\vec{E}_{\text{instantaneous}} = \text{Re}(e^{j\omega t}\ \vec{E}_{\text{complex}}) \qquad (11)$$

where ω is the angular frequency, $j = \sqrt{-1}$, and Re designates "Real part of." Now (7) reduces to

$$\nabla^2 \vec{E} + \frac{\omega^2}{c^2}\vec{E} = 0 \qquad (12)$$

which is the Helmholtz equation. The complex field representing waves of the type (10) is then

$$E_x = A\ e^{-j(\omega/c)z} + B\ e^{j(\omega/c)z} \qquad (13)$$

which, according to (11), represent waves

$$(E_x)_{\text{instantaneous}} = A \cos \omega(t - z/c) + B \cos \omega(t + z/c) \qquad (14)$$

By making A and B complex, an arbitrary phase can be included. The distance over which one cycle of the wave exists is called the wavelength

$$\lambda = 2\pi \frac{\omega}{c} \qquad (15)$$

and the commonly occuring parameter ω/c is called the wavenumber

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda} \qquad (16)$$

So far we have been considering only the simplest type of wave. Suppose we ask: What is the most general time-harmonic wave propagating (traveling) in one cartesian direction (say the z-direction) which is independent of the other two directions (x and y)? It is found that such a wave (called a uniform plane wave) must be transverse to z (the direction of propagation) and of the form

$$\vec{E} = (\vec{i}\, E_x + \vec{j}\, E_y)\, e^{-jkz} \tag{17}$$

If $E_y = 0$, then $\vec{E}_{\text{instantaneous}}$ points in the x direction at all times, and we have a wave that is linearly polarized in the x direction. If E_x and E_y are complex but have equal phases, the wave is still linearly polarized in some direction intermediate between x and y. If E_x and E_y are complex, equal in magnitude, and 90° out of phase, then lines of $\vec{E}_{\text{instantaneous}}$ rotate uniformly in space with angular frequency ω. Such a wave is said to be circularly polarized. More generally, a wave may rotate nonuniformly if E_x and E_y are not in phase, and such waves are said to be elliptically polarized.

If we consider waves traveling in opposite directions we obtain the phenomena of wave interference. The waves may add at some points in space and cancel at other points, forming a standing wave pattern. The simplest type of standing wave is that consisting of two equal but oppositely traveling waves

$$E_x = A(e^{-jkz} + e^{+jkz}) \tag{18}$$

which represents the instantaneous field

$$(E_x)_{\text{instantaneous}} = 2A \cos kz \cos \omega t \tag{19}$$

Such a wave has complete nulls at $kz = n\pi/2$, $n = \pm 1, \pm 3, \pm 5, \ldots$ Partial standing waves have minima and maxima stationary in space, but not complete cancellation.

Standing waves commonly occur whenever a wave is reflected by a material object. The simplest type of reflection is that of a uniform plane wave normally incident on a plane conductor. The reflected wave must be equal in amplitude to the incident wave, and of phase such that $E = 0$ at the conductor surface. Hence, for a conductor covering the $z = 0$ plane,

$$E_x = E_{\max} \sin kz \tag{20}$$

if the incident wave is x polarized. If the incident uniform plane wave travels at an angle Θ with respect to the direction perpendicular to the conductor, it is reflected at the same angle Θ to the other side of the normal. If the conductor is of finite size and irregular shape, the wave is scattered in many directions producing standing waves near the conductor. The solution in such a case is usually a difficult boundary-value problem.

If four plane conductors are arranged in the form of a hollow metal tube of rectangular cross section, a simple combination of two plane waves represents a possible field. With conductors covering the planes x = 0, x = a, y = 0, and y = b, this solution is

$$E_x = A \sin (\pi \frac{y}{b})\ e^{-j\beta z} \tag{21}$$

where

$$\beta = k \sqrt{1 - \frac{\pi c}{\omega b}} \tag{22}$$

is the phase constant. Equation (21) represents a wave propagating down the tube whenever β is real. According to (22), β is real only when

$$b > \frac{\pi c}{\omega} = \frac{\lambda}{2} \tag{23}$$

that is, when b is greater than a half wavelength. Assuming b > a, the above solution represents the dominant mode of the tube (called a waveguide), with propagation at lower frequencies than any other modes (called higher-order modes). Hence, a waveguide acts as a low-pass filter, propagating waves only when (23) is satisfied. Metal tubes of arbitrary cross section behave similarly, but present more difficult mathematical problems.

II. Radiation

So far we have talked of waves as existing, with no consideration of their sources. An electric current $\vec{J}$ and its associated charge ρ represent the source of an electromagnetic field. It enters into Maxwell's equations according to

$$\nabla \times \vec{B} = \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t} + \mu_o \vec{J} \qquad (24) \qquad\qquad \nabla \cdot \vec{E} = \rho/\epsilon_o \qquad (25)$$

with the other two, (1) and (4), being unchanged. Here ϵ_o and μ_o are the permittivity and permeability of vacuum, chosen as

$$\mu_o = 4\pi \times 10^{-7} \qquad\qquad \epsilon_o = \frac{1}{\mu_o c^2} \tag{26}$$

for the MKS system of units. The Helmholtz equation (12) now becomes the inhomogeneous equation

$$\nabla^2\vec{E} + k^2\vec{E} = j\omega\mu_o \vec{J} + \frac{1}{\epsilon_o}\nabla\rho \tag{27}$$

from which the complex $\vec{E}$ must be determined.

Rather than solve (27) directly, it is convenient to introduce a vector potential $\vec{A}$ such that

$$\vec{B} = \nabla \times \vec{A} \tag{28}$$

$$\vec{E} = -j\omega\vec{A} + \frac{c^2}{j\omega}\nabla(\nabla \cdot \vec{A}) \tag{29}$$

The equation for $\vec{A}$ is then found to be

$$\nabla^2\vec{A} + k^2\vec{A} = -\mu_o\vec{J} \tag{30}$$

which is simpler to treat than (27) because rectangular components of $\vec{A}$ depend only on the corresponding rectangular components of $\vec{J}$. For $\vec{J}$ in unbounded space, the solution to (30) is given by

$$\vec{A} = \int \mu_o\vec{J}\,\frac{e^{-jkr}}{4\pi r}\,d\tau \tag{31}$$

which is an integral over all $\vec{J}$. Hence, for any problem in which the electric currents are known everywhere in space, we have a solution given by (28), (29), and (31).

The simplest type of source is the current element, of strength I over an incremental distance ℓ. In this case the integrand of (31) is an impulse function, giving

$$\vec{A} = \frac{\mu_o\,\vec{I\ell}}{4\pi r}\,e^{-jkr} \tag{32}$$

The product $\vec{I\ell}$ is called the moment of the source. Equatioas (28) and (29) are most easily evaluated in spherical coordinates, chosen such that $\vec{I\ell}$ is at $r = 0$ and in the polar direction ($\theta = 0$). The result is then

$$E_r = \frac{\mu_o I\ell}{2\pi} \; e^{-jkr} \left(\frac{c}{r^2} + \frac{c^2}{j\omega r^3}\right) \cos \Theta \tag{33}$$

$$E_\Theta = \frac{\mu_o I\ell}{4\pi} \; e^{-jkr} \left(\frac{j\omega}{r} + \frac{c}{r^2} + \frac{c^2}{j\omega r^3}\right) \sin \Theta \tag{34}$$

$$B_\phi = \frac{\mu_o I\ell}{4\pi} \; e^{-jkr} \left(\frac{jk}{r} + \frac{1}{r^2}\right) \sin \Theta \tag{35}$$

This is a rather complicated result, but can be understood when viewed in parts. Very close to the current element, the $1/r^3$ terms dominate, and represent the electrostatic field from a charge dipole. The dominant $1/r^2$ term in B represents the magnetostatic field of a current element. The associated $1/r^2$ terms in E are the induction field of the current element, obtainable by neglecting the displacement current. At large r, the 1/r terms dominate, giving the radiation field

$$E_\Theta = \frac{j\omega\mu_o I\ell}{4\pi r} \; e^{-jkr} \quad \sin \Theta \tag{36}$$

$$B_\phi = \frac{jk\mu_o I\ell}{4\pi r} \; e^{-jkr} \quad \sin \Theta \tag{37}$$

This radiation is a maximum at right angles to $I\ell$ (at $\Theta = 90^\circ$), and has a null at $\Theta = 0$ and 180°. The field pattern (magnitude of E or B) is doughnut-shaped with $I\ell$ pointing in the direction of the nulls.

When the source $\mu_o I\ell$ is unity, from (32) we have

$$A_u = \frac{e^{-jkr}}{4\pi r} \tag{38}$$

This field of a unit source is called the Green's function for A, and the general solution (31) can be written as

$$\vec{A} = \int \mu_o \vec{J} A_u \, d\tau \tag{39}$$

The form of (39), which is a summation over all sources times a Green's function, is very general. The solution of any inhomogeneous differential equation may be put into such a form. The particular Green's function to be used depends on the medium into which the source radiates. Mathematically, this is equivalent to saying that it depends both on the form of the equation and on the boundary conditions. Once the solution to a unit source is found for any particular problem, then in principle the solution for any source is known. With this in mind, let us discuss some solutions for elemental sources in the vicinity of spheres.

In the spherical coordinate system, general solutions to the field equations in homogeneous media can be constructed from two scalar wave functions ψ according to

$$\vec{E} = \nabla \times \vec{r}\, \psi_1 + \nabla \times \nabla \times \vec{r}\, \psi_2 \qquad (40)$$

where r is the radius vector. The ψ's are linear superpositions of solutions to the scalar Helmholtz equation, which are of the form

$$\psi = z_n(kr)\, L_n^m\, (\cos \Theta)\, e^{jm\phi} \qquad (41)$$

The z_n represent spherical Bessel functions, and L_n^m solutions to the associated Legendre equation. The particular Bessel and Legendre functions chosen depend on the problem.

If the field external to a spherical surface is desired, then the ψ's must be finite and represent outward traveling waves at infinity. A study of Bessel and Legendre functions reveals that the only solutions of this type are

$$\psi = h_n^{(2)}(kr)\, P_n^m(\cos \Theta)\, e^{jm\phi} \qquad (42)$$

where n and m are integers, $h_n^{(2)}$ is the spherical Hankel function, and P_n^m is the associated Legendre polynomial. The ψ_1 and ψ_2 are linear combinations of (42), hence

$$\psi_1 = \sum_n \sum_m C_{nm}\, h_n^{(2)}(kr)\, P_n^m(\cos \Theta)\, e^{jm\phi} \qquad (43)$$

with a similar equation for ψ_2. The C_{nm} are determined from boundary conditions at the sphere and source.

One of the simplest problems to solve explicitly is that of a radially-directed current element $I\ell$ at the pole of a conducting sphere. In this case the radiation field is given by

$$E_\theta = \frac{I\ell\ e^{-jkr}}{4\pi j\omega\epsilon r} \sum_{n=1}^{\infty} \frac{j^n(2n+1)}{\hat{H}_n{}'(ka)} P_n{}^1(\cos\theta) \qquad (44)$$

where $\hat{H}_n(ka) = ka\ h_n^{(2)}(ka)$ is a spherical Hankel function. For a very small sphere, this is the field of a dipole. For a very large sphere, the pattern is a distorted doughnut, with the lobes pushed upward in the direction of the current element.

Equally important as a source of radiation as the current element is the aperture radiator. This consists basically of a hole in a conducting body through which electromagnetic energy can escape. A simple to analyze aperture problem is that of a conducting sphere sliced in half and excited by a voltage V across the resulting slot, assumed to be in the $\theta = 90^\circ$ plane. Once again we can take the general solution of (40) and (43), apply boundary conditions at the sphere $r = a$, and obtain the radiation field

$$E_\theta = \frac{jV\ e^{-jkr}}{2\pi r} \sum_{n=1}^{\infty} \frac{j^n(2n+1)\ P_n{}^1(0)}{n(n+1)\ \hat{H}_n^{(2)}{}'(ka)} P_n{}^1(\cos\theta) \qquad (45)$$

This field is that of a dipole when the sphere is small, and becomes almost omnidirectional for large spheres, except for severe interference phenomena in the polar regions and nulls in the axial directions.

Other problems that can be analyzed with the formulation of (40) with ψ's different from (42) are currents on and apertures in conducting cones. The problem is more difficult than the corresponding sphere problem because the solution involves Legendre functions of nonintegral order. Some examples will be given in the lecture.

III. Antennas and Arrays

Two important purposes of antennas are (a) to provide an efficient transfer of energy from a source to waves in space, or vice versa, and (b) to direct this energy in some desired manner. Parameters of interest for purpose (a) are the input impedance and efficiency of the antenna. Parameters of interest for purpose (b) are gain, beamwidth, and sidelobe level. Antenna arrays are groups of antenna elements connected together to provide desirable overall characteristics. The beamwidth of an antenna is the frequency range over which specified parameters remain within some desired range of variation.

Let us first define some of the above used terms. The gain G of a transmitting antenna is defined as

$$G = \frac{\text{Maximum density of radiated power}}{\text{Average density of radiated power}} \qquad (46)$$

The effective area A of a receiving antenna is defined as

$$A = \frac{\text{Power delivered to matched load}}{\text{Power density of incident wave}} \qquad (47)$$

If the antenna and surrounding media are reciprocal, then

$$A = \frac{\lambda^2}{4\pi} G \qquad (48)$$

When nonreciprocal media are present (for example, plasmas or ferrites) then G and A are independent quantities. The beamwidth of an antenna is the angle between half power points on the main beam. The sidelobe level of an antenna is the ratio of the intensity of the mainlobe peak to that of the maximum sidelobe.

Small antenna elements are basically of two types: (a) electric dipoles and (b) magnetic dipoles. Practical electric dipoles are formed by either a pair of conductors or a loop aperture. Practical magnetic dipoles are formed by a conducting loop or a small aperture. Both electric and magnetic dipoles have a gain G = 1.5 and doughnut-shaped patterns. With respect to input impedance they are basically narrow-band devices, since the impedance is a rapidly varying function of frequency. With respect to radiation pattern, they are not very directive, and relatively broadband (pattern not a function of frequency). Dipoles antennas are used when simplicity of construction and nondirective radiation are desired. They are also often used as elements of an antenna array.

Whenever an antenna is mounted on a body, the entire structure becomes a radiating system. For example, an electric dipole on a sphere or a cone has a very different radiation pattern than does the same dipole in free space. When designing antennas for spaee vehicles, it becomes very important to take the vehicle into account. Since the vehicle is usually of complex shape, one can use theory only as a guide to the design. The final design requires experimental measurements, either on models or on the actual structure. Because of the relatively small size of space vehicles, low frequency antennas cannot be highly directional, as discussed below.

Ground based antennas can be large, and usually are highly directional. The two commonest types are (a) antennas focused by a shaped reflector, and (b) arrays of many radiating elements. Type (a) has the advantage of design simplicity, and (b) the advantage of versatility. For very large antenna systems, the physical pointing of reflector antennas becomes difficult, and the ability to electrically steer an array becomes important.

Before discussing specific types of antennas further, let us see how some of the basic limitations to the behavior of antennas come about. Consider an arbitrary antenna and let "a" be the radius of the smallest sphere that can contain it. External to $r = a$ the field can be expanded in terms of spherical wave functions, according to (40) to (43). The actual antenna structure determines the coefficients C_{mn} of (43). We can determine the gain of the antenna according to (46) with the C_{mn} arbitrary, and inquire as to what is the maximum possible gain. The result is that, if all modes are allowable (all m and n), there is no limit to the gain. However, if we look at each mode of free space, we find that whenever $r < n\lambda/2\pi$ the m,n-th mode becomes cut-off, and cannot propagate energy effectively. Hence,• the summation of (43) is limited in practice to $n < ka$, and one finds a practical limit to the gain

$$G_{\text{normal}} = (ka)^2 + 2ka \tag{49}$$

called the <u>normal gain</u> of an antenna. Equation (49) is not an absolute theoretical limit, and we may inquire as to what happens if a higher gain antenna, called a <u>supergain</u> antenna, is designed. Careful analysis shows that supergain antennas rapidly (a) become very narrow band devices, (b) have high field intensities in the vicinity of the antenna structure, (c) have excessive power losses in the antenna structure, and (d) require extreme precision of design and construction.

So rapidly do these characteristics become evident that it appears that supergain antennas are impractical. Since reflector-type antennas commonly come within 80 percent of (49), and array antennas have a gain about equal to the normal gain, we can already do about as well as can be expected with convention designs, so far as gain is concerned.

By special techniques, it is possible to improve some antenna characteristics at the expense of others. For example, the beamwidth of a large receiving array can be obtained by multiplying the output of two linear arrays at right angles to each other (called a Mills' cross), if only a single signal is received. Some disadvantages of this design are (a) lower gain that the larger array, (b) false indications if two signals are present, and (c) it can be used only for receiving purposes. For radar purposes, the effective beamwidth of an array can be reduced in theory by using more than one frequency of transmission. A movable antenna can give the effect of a larger array by using a signal storage system. Signal processing schemes can be used to treat the output of each element of an array as a separate signal, thereby enhancing some antenna characteristics at the expense of others. However, none of these schemes can circumvent the fundamental limitation (49) on gain.

References:

1. H. H. Skilling, "Fundamentals of Electric Waves," John Wiley and Sons, Inc., New York, 1948 (introductory level).

2. Ramo and Whinnery, "Fields and Waves in Modern Radio," John Wiley and Sons, New York, 1953 (intermediate level).

3. E. C. Jordan, "Electromagnetic Waves and Radiating Systems," Prentice-Hall, Inc., New Jersey, 1950 (intermediate level).

4. R. F. Harrington, "Time-Harmonic Electromagnetic Fields," McGraw-Hill Book Co., New York, 1961 (advanced level).

5. J. D. Kraus, "Antennas," McGraw-Hill Book Co., New York, 1950 (introductory level).

6. R. F. Harrington, "Effect of Antenna Size on Gain, Bandwidth, and Efficiency," Journal of Research N. B. S., vol. 64D, No. 1, Jan. 1960.

7. A. A. Ksienski, "Signal Processing Antennas," Microwave Journal, vol. 4, Nos. 3 and 4, Oct. and Nov. 1961 (survey article).

1: Coding, Filtering, and Information Theory

by Leonard S. Schwartz In ... p 15-35 refs
(See ...)

1. Coding Principles

Our theory of communications as a statistical problem is largely the work of two men, Claude Shannon and Norbert Wiener. Shannon's theory basically is one of coding and shows how to organize a message at the transmitter--within certain limits of power, bandwidth and time--in order to resist the corrupting effect of channel noise. Wiener's theory basically is one of filtering and shows how to recover--within limits imposed on the same parameters--a noise-corrupted signal at the receiver.

As we can see, transmitter power, signal bandwidth, and message duration are key factors in determining communication rate. R. V. L. Hartley, in 1928, first formulated the basic relationship of these parameters by taking the amplitudes of signal waveforms as quantized in N levels, with each level representing a possible state of the waveform source (Fig. 1). In an interval of T seconds and with a signal bandwidth of W CPS, there are 2WT independent samples. Within N discrete levels there are thus N^{2WT} possible states. The information contributed by selection of the various states is proportional to the log of N, and the information rate is:

$$R = 2WT \log N. \tag{1}$$

However, Hartley's theory failed to consider the fineness of quantization, omitted any reference to noise, and did not take into account the probabilities of the various states of the message source. His concept of a source with an output unperturbed by noise therefore has been replaced by Shannon's concept of a statistical source connected to a noisy channel.

For transmission, information is measured by the number of different messages the transmitter is theoretically able to transmit. At the receiver, information is measured in terms of the initial uncertainty about which of these messages will be transmitted. Reception of the message removes this uncertainty. The greater the number of messages available at the transmitter, the greater the initial receiver uncertainty, but the greater the information transferred when the message arrives at the receiver. Final information is therefore a function of message probability.

The measure of initial uncertainty is a quantity analogous to entropy in statistical mechanics. Entropy is a measure of randomness or disorder, and is expressed in terms of the logarithm of probabilities. In the case of communications, entropy may be regarded as negative information, or the initial uncertainty at the receiver that is removed when the message is received. Information therefore can be defined as:

$$h(p) = - \log_2 p, \tag{2}$$

where p is the message probability. The logarithmic base b may have any arbitrarily chosen value, but the base 2 is usually simplest to handle. The basic unit of information is the bit, or binary digit, which results from the selection between two equally probable alternatives.

A source has maximum entropy when all symbols are equally likely as well as all their conceivable sequential arrangements. Otherwise, the source is redundant. Redundancy is defined as the amount by which the logarithm of the number of available symbols at the source exceeds the average information content per symbol.

A redundant source uses more symbols than absolutely necessary to transmit a given amount of information, but it is useful because it reduces the probability of error in the presence of noise. Our language, for example, is highly redundant. A single mistyped letter in an English text can usually be corrected without trouble--the information content of the missing correct letter is duplicated by other letters and their arrangement.

The link between the transmitter and receiver is the communication channel, which includes the transmission medium and usually the transmitter and receiver modulation and demodulation equipment (Fig. 2). For maximum information transfer from message source to destination, source and channel must be statistically matched. The statistical nature of the messages is determined entirely by the source properties; but the statistical nature of the channel and therefore the entropy in the channel are determined by what is transmitted and by the ability of the channel to receive different signals.

An Important Expression Was Found

The channel capacity (C) is the maximum rate at which information can be sent. Since channel symbols and transition probabilities are known, maximazation requires only the best choice of input probabilities. For the case of noise acting independently on successive symbols, Shannon found an important expression for the capacity of a continuous channel perturbed by thermal noise:

$$C = W \log_2 (1 + P/N) \text{ bits per second,} \qquad (3)$$

where W is the channel bandwidth; N, the thermal noise power; and P, the average signal power.

After sufficiently involved encoding, binary digits can be transmitted at the maximum rate (C) with an arbitrarily small frequency of errors. Regardless of the encoding system, transmission at a higher rate is impossible without a definite error frequency.

In general, the theoretical limit of C can be approached only if complex encoding and very long messages are used. Greater message lengths do not imply lower information rates, for information is received throughout the period of message reception. Moreover, the longer the message, the easier it is for the receiver to go back and correct errors in parts of the message that have already been received.

In a noise-free channel, a continuous function of time ideally can be measured with high precision and can assume an infinite number of amplitude levels. An unlimited amount of information can therefore be transmitted per unit time over as narror a bandwidth as desired. In practice, bandwidth can be traded for signal-to-noise ratio. Also, a message can be transmitted over a channel with a smaller bandwidth than the original signal's by changing this signal to a new form, or coding it. This can be done so long as the power has the magnitude required by Equation 3.

The performance of practical communications systems lies well below the theoretical upper limit given by Equation 3. PCM, for example, one of today's most effective systems, requires about eight decibels more than the theoretical signal power for a given channel capacity at a given bandwidth.

As we have noted, redundancy requires errors in received messages, but it still makes for inefficient operation. If the S/N is high enough to make errors unlikely, the redundancy and therefore the system bandwidth may be reduced. As an example, you may want to transmit English speech and require only that the received speech signals be intelligible--personal accents, inflections, and the like may be lost. In this case, a text corresponding to spoken words is printed at the transmitter and encoded into binary digits. On the average, not more than two digits will be used per letter, or nine per word. Strictly on the basis of intelligibility, a 100-wpm speaking rate then corresponds to an information rate of 15 bits per second. If the received S/N ratio is 40 db, a bandwidth of about one cycle will do (Equation 3).

Theorem Defines Reliability Limits

An important goal of communications design is to maximize the rate of reliable communication. The limits on the maximum rate of information transfer over a noisy channel and also on the reliability of the received information are defined by Shannon's noisy-channel coding theorem. It states that a transmitted message can be received with an arbitrarily small error probability if the transmission rate (R) is less than channel capacity (C). Conversely, the error probability cannot be made arbitrarily small if R equals or exceeds C. R and C are independent of each other, since C is a channel characteristic with the same value for all receiver-transmitter pairs, while R is a transmitter-receiver-pair characteristic with the same value for all channels.

Ideally, error-free information can be sent at some finite rate which is arbitrarily close to the channel capacity. A non-zero information rate for a vanishingly small error probability can be achieved by systematic coding. Although this rate is definitely less than the channel capacity, it represents a remarkable achievement--less than 20 years ago, it was considered impossible.

Coding that gives better reliability is an error-correcting process. R. W. Hamming introduced the codes of this type by defining a systematic-block, or parity-check, code as one in which each character has a block of n binary digits, of which k carry information. The remaining n - k digits are check, or parity, digits for error detection and correction. When n - k = 1, single errors can be detected by using the check digit to give an even number of ones for each transmitted character. If the count for any received character is odd, an error is present. However, the location of an error within a character cannot be found by this technique alone.

In iterative systematic codes, decoding is done over short blocks of digits, and the corrected digits are used together with other blocks for further error correction. In sequential codes, a small number of digits is decoded at a time. For both iterative and sequential coding, the error probability decreases exponentially with code length, while the number of computations goes up only as a constant power of length.

Rate Can be Close to Capacity

Optimum coding (infinite delay) can provide an error-free rate close to channel capacity, but the same rate results with simple constructive coding (finite delay) if noise-free feedback is added. Noise in the feedback path merely increases the delay required to achieve the maximum rate.

The forward-channel capacity remains unchanged by the presence of the feedback path, unless the noise in the two channels is correlated (increasing or decreasing at the same time), in which case the forward-channel capacity may be increased (Fig. 3). Feedback has shown marked advantages over coded single-channel operation, but even better results are achieved if the two techniques are combined.

2. Filtering Principles

Until the forties, no methods existed for analyzing random functions, which play a major role in communications because of the random nature of signals and noise. Wiener's development of generalized harmonic analysis, of which repetitive, or aperiodic-function, analysis is a special case, provided communications engineers with mathematical tools they had long been waiting for.

Random functions are characterized by the fact that repeated experiments under similar conditions produce results that are of the same general type but act instantaneously identical. An example would be a set of fading records of radar signals from an "ensemble" of aircraft in all possible states of motion. Each record, made for all the planes over a long period of time, would be a random time function. If all the fading records were placed parellel in time (so that all T = 0's coincide), they would form an ensemble of random time functions, or a random process.

In many practical situations, the statistical properties of the ensemble are invarient with a shift in time, forming a stationary random process. This process is described by a set of probability distributed functions that are difficult, if not impossible, to determine either analytically or experimentally.

For a number of communication problems, the so-called second probability distribution function adequately describes the random process. Most useful, however, is the autocorrelation function--though dependent on the second distribution, it effectively characterizes the random process. This easily calculated function is expressed as an ensemble average:

$$\psi(\tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1 y_2; \tau) dy_1 dy_2, \tag{4}$$

where y_1 and y_2 are the amplitudes of time functions taken τ seconds apart on the given waveforms, and $f(y_1, y_2; \ )$ is the joint probability density function for the process at points τ seconds apart. The autocorrelation function also may be expressed as time average (ergodic hypothesis):

$$\Psi_{11}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_1(t) f_1(t + \tau) dt. \tag{5}$$

Wiener's theorem gives the counterpart of the autocorrelation function in the frequency domain as:

$$\Psi_{11}(\tau) = \int_{-\infty}^{\infty} S(f) \exp(j\omega\tau) df, \tag{6}$$

and the power spectral density as:

$$S(f) = \int_{-\infty}^{\infty} \Psi 11(\tau)\exp(-j\omega\tau). \tag{7}$$

The autocorrelation function and the power spectral density are Fourier transforms of each other. This reciprocal relationship characterizes the stationary random process--if it is valid for any single member of the statistical ensemble, it is valid for all. The situation is analogous to linear system behavior, which may be expressed as a transfer function (frequency domain) or as a response to a unit impulse function (time domain). Each function is the Fourier transform of the other, and each brings out certain information in a better light (Fig. 4).

A similar relationship exists between two stationary random processes that are related to each other in some manner. If $\Psi_{12}(\tau)$ is the cross-correlation function of two coherent stationary random processes defined by:

$$\Psi 12(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_1(t)f_2(t + \tau)dt. \tag{8}$$

where $f_1(t)$ and $f_2(t)$ are member functions of the two processes, then we can write:

$$\Psi 12(\tau) = \int_{-\infty}^{\infty} S_{12}(f)\exp(j\omega\tau)df, \tag{9}$$

$$S_{12}(f) = \int_{-\infty}^{\infty} \Psi_{12}(\tau)\exp(-j\omega\tau)d\tau. \tag{10}$$

where $S_{12}(f)$ is the cross-spectral power-density function.

Correlation Relates a Signal's Parts

Correlation generally expresses relationships between parts of the same signal or parts of several signals. Different parts of a completely random signal are independent of each other. In this case, the value of the autocorrelation function is zero (except when τ equals zero), signifying that the signal is correlated with itself only. The autocorrelation function of thermal noise, for example, is a spike at $\tau = 0$. When some correlation is present, the autocorrelation function decays to zero with increasing τ, except for periodic functions.

The cross-correlation function is a measure of the correlation between two different function--it gives the average product relationship between them. Cross-correlation is useful primarily when a priori information is available about the signal frequency.

In accordance with Wiener's theory of linear filters, correlation can be used to recover non-random signals from a background of random noise. The linear filters generally are regarded as correlators or integrators in the time domain or as matched and "comb" filters in the frequency domain. For example, random processes may have hidden periodic components that, in some applications, must be separated from the whole. Autocorrelation provides this separation, since the autocorrelation function of a periodic signal is periodic at the same frequency. Random-function autocorrelation, on the other hand, has a maximum equal to the mean square value when τ equals zero and falls off to the square of the mean value at $\tau = \infty$. Cross-correlation offers an advantage if the signal period is known and particularly if the S/N is much less than unity.

As correlation achieves the same results in the time domain that an extremely narrow-band filter achieves in the frequency domain, it sometimes can be used with good effect to obtain a signal spectrum. A case in point is the isolation of a sine-wave signal hidden by random noise.

Correlation is used in synchronous detection for color TV. A message function $f_m(t)$ is modulated by a cosine (or sine) voltage at the transmitter to give a signal $f_2(t)$:

$$f_m(t) \cos \omega_o t = f_2(t). \qquad (11)$$

The received signal is multiplied by a cosine (or sine) reference voltage of the carrier frequency to give:

$$f_2(t) \cos \omega_o t = f_m(t) \cos^2 \omega_o t$$

$$= \frac{f_m(t)}{2} (1 + \cos^2 \omega_o t). \qquad (12)$$

The signal then passes through a low-pass filter, which eliminates the carrier-frequency, and the output is $f_m(t)/2$.

To the extent that synchronous detection relies on multiplication with a known frequency and subsequent averaging, it is cross-correlation. But it is incomplete correlation--the band of harmonic components constituting the signal is passed by the filter, while true correlation functions as an infinitely narrow-band filter and passes the carrier frequency only.

3. Feedback Principles

The purpose of several recent developments in the field of digital communications is reliability improvement, particularly as related to the method of decision in the receiver and introduction of feedback channel between the receiver and transmitter.

It is not suggested that these methods take the place of coding, or even that they are in any way superior. Rather, they represent additional means for coping with the reliability problem in communications, and it is hoped that they may be combined with coding to effect greater improvements than could be achieved with coding alone.

Null Detection

The technique of null detection[1] permits a decision to be withheld in doubtful cases and new observations to be taken until unambiguous readings are obtains. Transmission need not be fixed in advance, since it may be determined during the course of communication by criteria that depend on the received signals. The method is, on the average, more efficient in saving time for a given reliability or increasing the reliability for a given time than an optimized, single-threshold system, a result that is intuitively reasonable because null detection reduces guessing, which destroys information.

The null method is implemented by introducing two threshold levels at the receiver, thus permitting the receiver to withhold decision in doubtful cases, so that a large reduction in decision error probability is possible, but would be accomplished at the expense of producing null or rejected signals. If the transmitted message contains either the natural redundancy of language or the artificial redundancy of code, some of the null signals may be interpreted correctly. The great advantage of null detection consists in the pinpointing of the most probable errors, whereas in ordinary binary transmission the positions of errors must first be located before corrections can be made.

When the input information consists of a statistically long sequence of uncoded binary symmetric digits and the interference is symmetrical, the channel may be characterized by the probabilities p_i of receiving a digit incorrectly, u_i of receiving a digit ambiguously or as a null, and q_i of receiving a digit correctly, where the subscript $i(1 = i = \infty)$ denotes the values of the probabilities on the fth transmission. The decision process at the receiver places the received signal y_i in one of three groups:

Group I	$y_i \geq k_i$
Group II	$- k_i < y_i < k_i$
Group III	$y_i \leq - k_i$

where k_i is the voltage level of the null boundary on the fth transmission.

The following example for a binary-to-ternary channel illustrates the effects of nulls in accepting or rejecting coded messages. The channel transition diagram is shown in Fig. 5. We shall assume coding according to a system of even parity checks. Suppose the transmit-receive message pair is

$$1011\underline{1} \longrightarrow 10\phi\ 1\underline{1}$$

In this case the receiver infers that most probably the null symbol (ϕ) should be replaced by a "1" because this would satisfy an even parity check. It is recognized that the presence of the null zone does not remove all errors. It only reduces their probability, since some symbols that otherwise would be changed to their opposites become nulls instead.

Moreover, the message rejection probability is reduced, because the message would be rejected in a binary symmetric channel can be corrected and accepted in a 2-3 channel, provided only a single null has occurred. If more than a single null has occurred or if a null and an error have occurred, the message must be rejected.

Limitation of decision to three levels--0, 1, and --may destroy some information, because knowledge of the detailed features of the noise structure may thereby be ignored. Thus, in the case of Gaussian noise, an appreciable improvement in per-symbol information rate may be achieved by increasing the number of decision thresholds in the receiver. With an infinite number of thresholds it would be possible for the receiver to preserve all information in making a decision.

A receiver operating on the decision plan just described is a single-null receiver. It has been found that if the number of thresholds in the receiver is increased without limit, the improvement in reliability that results corresponds to a gain of about 2 db in signal-to-noise ratio over a single threshold system. With just two thresholds (single null) the improvement corresponds to a gain of about 1 db. Double-null detection (four decision thresholds) results in further improvement in reliability, but only to a small degree. Because of the rapid growth in the complexity of the decision mechanism with the number of decision thresholds and the correspondingly smaller degree of reliability improvement, discussion is usually limited to the case of two decision thresholds--that is, to a single-null receiver, henceforth called a null receiver for simplicity.

The null-detection technique, however, is valuable primarily when used in conjunction with feedback because the aim is not to lose the information in the rejected message, if possible. The only way this loss can be prevented is to be able to request the transmitter to repeat the rejected message until it can be accepted. The repetition is accomplished by use of a return path from the receiver to the transmitter; that is, feedback.

Kinds of Feedback

The addition of a feedback channel gives greater flexibility to the problem of error control in the forward channel. Specifically, it renders possible the replacement of nulls by valid 0's and 1's, which may effectively increase the information rate for a given reliability or increase the reliability for a given information rate on the forward channel.

Broadly speaking, feedback systems may be classified according to the type of communication carried by the feedback channel[2]. They are:

1. Decision feedback, in which the feedback channel is used only to report the decision of the receiver as to acceptability of the received message.

2. Information feedback, in which the feedback channel is employed to report information about the received message to the transmitter, with the decision to accept, or reject and correct, being made subsequently at the transmitter.

Decision Feedback

The simplest feedback system is a discarding decision system in which the receiver, as a result of a decision made on each received message, either accepts the message and records the corresponding symbol or rejects it as ambiguous and reports the rejection to the transmitter. The information in the ambiguous message is discarded, and the transmitter subsequently repeats the message. Adjustment of the null zone thus permits a trade-off between information rate and error probability[3].

Most of the analysis to date has assumed the feedback channel to be error-free. When this is the case, the information rate and the error probability of the direct transmission channel remain unaltered if the feedback loop is opened. The important advantage of the feedback loop is that the entire message can eventually be received at the specified error probability. The assumption that the errors in feedback can be neglected is reasonable in some cases. Thus, binary transmissions are usually composed of basic code groups, representing the symbols of some message alphabet. If only YES-NO information need be transmitted over the feedback channel for each code group in the forward channel, the effective SNR (signal-to-noise ratio) in the feedback channel can usually be maintained at a considerably higher value than in the forward channel. For most types of noise, error probability decreases quite rapidly with increasing SNR, and the error rate of the return channel can usually be kept small compared with that in the direct channel. Moreover, if the probability of acceptance is appreciably different from the probability of rejection, coding of the null locations will result in a considerable reduction of the required feedback capacity. The more precisely the probability of rejection is known, the more efficient this coding can be.

The possibility of combatting so-called fast fading of scatter-multipath reception by means of feedback has been studied[4]. It was assumed that the SNR has a Rayleigh distribution and that the interference is additive Gaussian. The system of prime interest in these investigations is one based on variable null-level reception and discarding decision feedback. When an effort is made to achieve a large reduction in error probability at the cost of increased transmission time, the inherent inefficiency of the discarding mechanism becomes a factor. Hence, a repetition--integration scheme has been considered, by means of which the feedback channel is used to monitor the SNR and consequently to optimize the number of repetitions. In this system, which has been called cumulative-decision feedback[4], each message digit is repeated continuously by the transmitter until the receiver indicates, via the feedback path, that the total received information is no longer excessively ambiguous. Each sequence of repeated digits is integrated in the receiver until the resultant signal passes through one of the two boundaries of the null zone, at which time the appropriate decision is signaled to the transmitter. As soon as a decision has been reached, the receiver returns the decision device to its central starting level, and transmission of another message digit is begun.

Cumulative-decision feedback yields, for the same guaranteed maximum error probability, an asymptotic rate four times as great as that attainable with a unidirectional variable repetition-integration system. The superiority of the cumulative scheme is attributed to the fact that it permits the transmitter to be responsive to the actual pulse-by-pulse needs of the receiver, rather than responsive in a statistical sense only.

One of the standard techniques for improving multipath reception is that of diversity reception. It is of interest, therefore, to ascertain the performance of the cumulative system just described when used in conjunction with dual diversity reception. The type of diversity reception considered is that in which the output of the channel having the better SNR is selected. It is found that a one-third additional reduction in average transmission time per information bit may be achieved as a result of adding the cumulative feedback feature.

Information Feedback

In information feedback systems[2,5] the transmitter has to provide additional capacity (erasure capacity) in order to keep the receiver informed as to whether it is conveying new information or corrective information. On the other hand, the receiver gains additional information by knowing whether or not the original message is acceptable to the transmitter. It can be shown that when feedback is error-free the gain in information at the receiver is equal to the net information of the confirmation--denial or erasure process. Thus, in these circumstances the proper use of an erasure signal does not cause a reduction in information rate.

The significant feature of information feedback, however, is not conservation of rate, but rather the possibility of transmitting error-free messages without coding. Error-free operation is possible if the feedback channel is error-free and of sufficient capacity to permit the receiver to inform the transmitter of the exact signal received via the direct channel. The transmitter can then, by means of the erasure process, correct all errors in transmission. The assumption of error-free feedback will be valid less often for information-feedback systems than for decision-feedback systems, because of the higher rates usually required of the feedback channel.

Loss of information in iterative-discarding-information feedback, in which incorrect information is erased without attempting to use it in any way, is assignable to three causes:

1. Noisy feedback
2. Rejection of residual information in erased words
3. Loss of information in the total signal entropy caused by the presence of the erasure word

The transition diagram in Fig. 6 will help to make clear the nature of the erasure problem which is peculiar to information feedback. It is assumed that any one of three symbols--0, 1, and θ (erasure symbol)--may be transmitted and any one of the three may be received. P stands for the previously received symbol.

Information feedback is highly effective, significantly more so than decision feedback, provided that the signal-to-noise ratio on the forward channel is of the order of 3 db or greater, although the improvement varies with the code length. The reason for the variation in performance is that if a coded forward transmission is rejected--that is, followed by an erasure symbol--the entire message must be repeated. It is apparent that the smaller the SNR the more frequently the message will be rejected. Moreover, the longer the code length, the greater the reduction in the resulting rate for a given SNR as a result of the rejection and repetition process. Thus, information feedback is characterized by a very sharp SNR threshold below which the information rate becomes zero for a given code length. On the other hand, decision feedback is always better for small values of SNR.

The excellent performance of the information-feedback system above the threshold is primarily attributable to control of the entire feedback process by the transmitter. It is noted that as the number of transmissions is increased the storage requirements must be increased in order to perform the iterative discarding process.

One method of avoiding the threshold effect is to change the coding scheme as the SNR changes as, for example, by a suitable amount of repetition-integration of each digit.

Coding Plus Feedback

A communication system that employs long codes, attempts error correction only for very small numbers of errors, and requires much less computation and storage than a system attempting the maximum possible error correction is the "long-code feedback system"[6]. Suitable operation of the feedback channel, which reports accept-reject decision to the transmitter, minimizes the effect of errors due to fading and disturbance in the feedback channel, a result that may be accomplished by utilizing a sufficiently asymmetrical decision mechanism at the transmitter to interpret the feedback information, so that reject-to-accept (R → A) errors occur with minimal frequency. The system possesses excellent reliability, fails safe even when the SNR drops to zero, and is particularly effective when less efficient procedures fail; that is, when severe burst-type noise or heavy fading is encountered. The cost of this performance in terms of computational and storage requirements is much less than that of comparable unidirectional systems. The outstanding characteristic of this system is its fail-safe feature under extreme variations in kind and extent of interference.

The coding procedure is to use an (n,k) group alphabet, also called a systematic or a parity check code, in which k digits of a sequence of n digits are information-carrying and c = (n - k) digits are check digits.

When the order of error correction is e = 1, n = 100, and c = 40, the error probability is 10^{-10}, which, at a bit rate of 1,000 bits per second, is equivalent to an average error rate of about one error every 16 years.

Two-Way Communication

The foregoing discussion has dealt with the case of accept-reject information returned to the transmitter by means of a signal constructed solely for this purpose. When message information is flowing in only one direction, some such arrangement is necessary. If information is to be transmitted in both directions, however, feedback can be used in both directions, and the proportion of the channel taken up by accept-reject information can be made negligibly small[7]. Thus, if A and B represent the two stations, (n_1, k_1) code groups with e_1 error correction can be used to transmit from A to B, and (n_2, k_2) code groups, with e_2 error correction from B to A. Only one digit in each group is needed to carry the accept-reject information regarding the appropriate code group received in the other channel. Each time A rejects a message, A sends a repeat of its own previous message along with a request for repetition of the previous B-to-A message. If, on the other hand, a message is accepted by A, there is almost no chance of error, and thus A may send either a new message or a repeat, depending on the received accept-reject digit. These remarks apply particularly if two-way information flow is desired over a single channel or over two channels so situated that signal and noise statistics in the two channels are somewhat dependent. If signal and noise statistics in the two channels vary independently, it may be desirable to separate the accept-reject information from the message-bearing groups.

Feedback Systems in Use

Rudimentary forms of feedback are not new to the communication art. One feedback system has been in operation for more than twenty years[8,9]. The ARQ (automatic request for repetition) system, proposed by Van Duuren and used extensively by RCA for long-range teletypewriter service, is a fixed-redundancy, discarding-decision feedback system. It employs a 3-out-of-7 constant-ratio code yielding 35 different code characters. Information flow is two way, one of the 35 allowable characters being reserved for requesting repeats. The system is intended for use under fairly good channel conditions.

Another form of ARQ system used on the Argentine teletype over radio (TOR) circuits, uses a form of digit-null reception also proposed by Van Duuren[10]. In this system the standard five-digit characters are transmitted over frequency-shift keying links; the form of the received signals is checked by testing them against a null zone. Whenever the tester rejects one element, the four preceding elements are also rejected and a repeat is requested to reduce the probability of erroneous acceptance during periods of deep signal fading. This system provides a fail-safe feature against certain types of disturbances, such as signal fading at constant noise level (10^{-5} to 10^{-7}) error probability. The system is less effective against bursts of noise and certain other types of interference that are likely to cause digit errors without causing a null to appear in any number of successive digits.

References

1. F. J. Bloom, et al.,"Improvement of Binary Transmission by Null-Zone Reception," IRE Proceedings, vol. 45, July 1957, pp. 963-75.

2. L. S. Schwartz, et al.,"Binary Communication Feedback Systems,"AIEE Transactions, pt. I (Communication and Electronics), vol. 77, 1958 (Jan. 1959 section), pp. 960-69.

3. B. Harris, et al., "Optimum Decision Feedback Systems," IRE National Convention Record, vol. 5, pt. 2, 1957, pp. 3-10.

4. S. S. L. Chang, et al., "Cumulative Binary Decision-Feedback Systems," Proceedings, National Electronics Conference, Chicago, Ill., 1958.

5. S. S. L. Chang,"Theory of Information Feedback Systems," IRE Transactions on Information Theory, vol. IT-2, 1956, pp. 29-40.

6. J. J. Metzner, K. C. Morgan,"Reliable Fail-Safe Binary Communication", IRE Wescon Convention Record, vol. 4, pt. 5, 1960, pp. 192-206.

7. S. S. L. Chang,"Improvement of Two-Way Communication by Means of Feedback", IRE International Convention Record, vol. 9, pt. 4, Mar. 1961, pp. 88-104.

8. A. Bakker, H. C. A. Van Duuren,"Type Printing Telegraph Systems with Means for Eliminating Interference", U. S. Patent no. 2,119,196, May 1958.

9. J. B. Moore,"Constant-Ratio Code and Automatic-RQ on Transoceanic HF Radio Services", IRE Transactions on Communication Systems, vol. CS-8, no. 1, 1959, pp. 72-75.

10. W. Six,"The T.O.R. Circuits for the Argentine Radio Links," Communication News, Camden, N. J., vol. 15, no. 4, 1955, pp. 108-19.

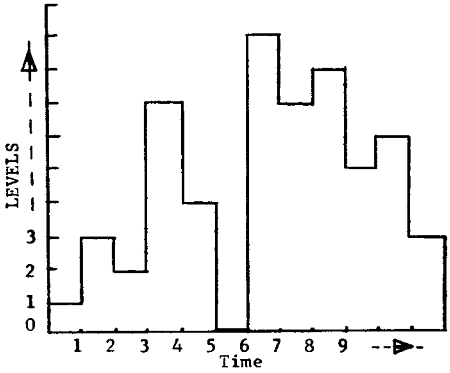


FIGURE 1: Quantization of signal-waveform amplitudes into N possible states of levels.

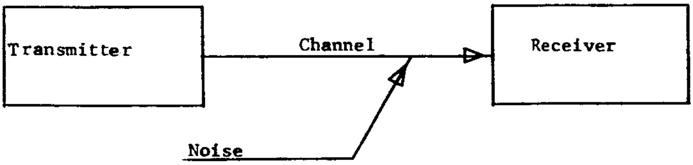


FIGURE 2: Basic elements of a communications system.

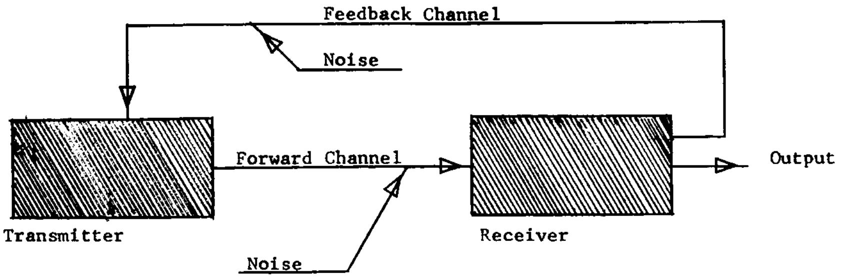


FIGURE 3: Basic communications system with feedback channel. Ideally, this channel is noise-free, and simple coding will bring the channel capacity close to the theoretical maximum.

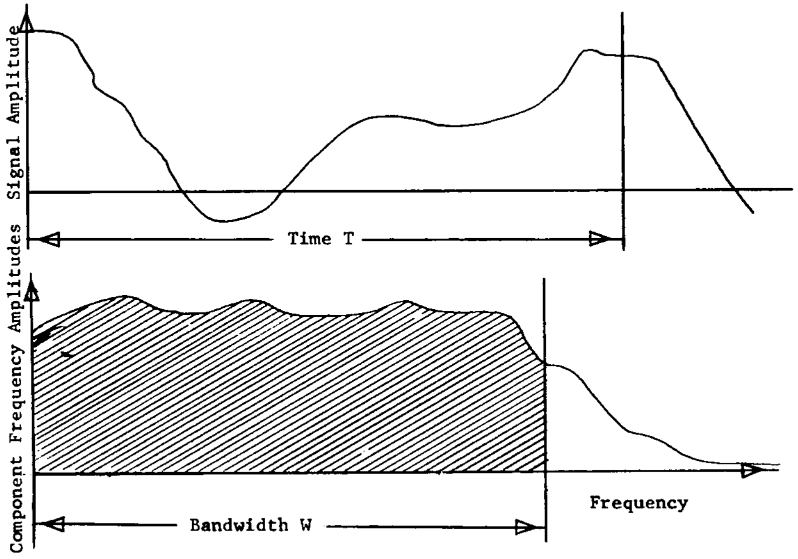


FIGURE 4: System behavior expressed as a function of time (top) and a function of frequency (bottom).

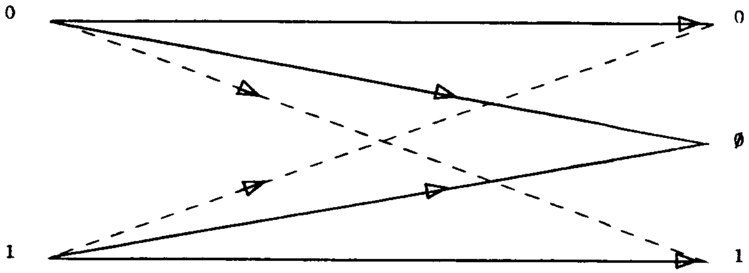


FIGURE 5: Transition diagram for 2-3 channel with null-zone decision system in the receiver.

R →

T ↓		0	1	θ
	0	-	θ0	P0
	1	θ0		P1
	θ	θθ	θθ	-

FIGURE 6: Transition diagram for iterative-discarding-information feedback.

17197

Alabama U. of Huntsville Research Inst.

8/3/672

t: PROBLEMS OF WAVE PROPAGATION

by F. J. Tischer In ... p 36-50 refs

(See ...)

I. Waves in an Infinite, Source-Free Region of a Uniform Medium (Review).

Maxwell's Equations:

$$\nabla \times \overline{E} = - \frac{\partial \overline{B}}{\partial t} ; \quad \nabla \cdot \overline{D} = 0$$

$$\nabla \times \overline{H} = + \frac{\partial \overline{D}}{\partial t} ; \quad \nabla \cdot \overline{B} = 0$$

$$\overline{D} = \epsilon \overline{E} \quad ; \quad \epsilon = \epsilon_o \epsilon_r$$

$$\overline{B} = \mu \overline{H} \quad ; \quad \mu = \mu_o \mu_r$$

Considering polarization vectors

$$\overline{D} = \epsilon_o \overline{E} + \overline{P} \quad ; \quad \overline{P} = \epsilon_o (\epsilon_r - 1) \overline{E}$$

$$\overline{B} = \mu_o \overline{H} + \mu_o \overline{M} \quad ; \quad \overline{M} = (\mu_r - 1) \overline{H}$$

Combination and evaluation leads to wave equations

$$\nabla \times \nabla \times \overline{E} + \epsilon \mu \frac{\partial^2 \overline{E}}{\partial t^2} = 0$$

$$\nabla \times \nabla \times \overline{H} + \epsilon \mu \frac{\partial^2 \overline{H}}{\partial t^2} = 0$$

For some coordinate systems these can be reduced to

$$\nabla^2 \overline{E} - c \mu \frac{\partial^2 \overline{E}}{\partial t^2} = 0$$

$$\nabla^2 \overline{H} - \epsilon \mu \frac{\partial^2 \overline{H}}{\partial t^2} = 0$$

by the vector identity

$$\nabla \times \nabla \times \overline{A} = \nabla (\nabla \cdot \overline{A}) - \nabla^2 \overline{A}$$

Simplest case for rectangular coordinates (x, y, z) (See Fig. 1)

$E_y = E_z = 0$:

$$\frac{\partial^2 E_x}{\partial z^2} - \epsilon\mu \frac{\partial^2 E_x}{\partial t^2} = 0$$

$$E_x = F_1 (t - z/v) + F_2 (t + z/v) \; ; \; v = 1/\sqrt{\epsilon\mu}$$

The two terms represent wave functions in the positive and negative z-direction. We find for the magnetic field intensity

$$H_y = \frac{F_1}{Z} - \frac{F_2}{Z} \; ; \; Z = \sqrt{\mu/\epsilon}$$

Free space:

$$V_o = 1/\sqrt{\epsilon_o \mu_o} = 3 \times 10^8 \text{ meter/sec}$$

$$Z_o = \sqrt{\mu_o/\epsilon_o} \approx 377 \text{ ohms}$$

II. Waves in Various Coordinate Systems for Harmonically Varying Sources (Review).

(Suggested references: Stratton [1], Harrington [2])

General elementary wave functions

a. Rectangular coordinates

$$\phi_1 = e^{-j\Gamma_x x} e^{-j\Gamma_y y} e^{-j\Gamma_z z} e^{jwt}$$

$$\phi_2 = e^{+j\Gamma_x x} e^{+j\Gamma_y y} e^{+j\Gamma_z z} e^{jwt}$$

$$\Gamma_x^2 + \Gamma_y^2 + \Gamma_z^2 = \Gamma_o^2 = \beta_o^2$$

$$\phi_{1,2} = e^{\mp j \overline{\Gamma} \cdot \overline{r}} e^{jwt}$$

$$j\Gamma = \gamma = \alpha + j\beta$$

γ, Γ, α, β -------field-distribution, wave-propagation, attenuation, and phase constants

Standing waves

$$\phi\ \mathrm{sl},\ 2 \ = \begin{bmatrix} \cos\ \Gamma_x x \\ \\ \sin\ \Gamma_x x \end{bmatrix} e^{jwt}$$

b. Cylindrical coordinates

Solution of the wave equation

$$\phi\ 1,2 \ = \begin{bmatrix} H_n^{(1)}\ (kr) \\ \\ H_n^{(2)}\ (kr) \end{bmatrix} e^{jn\phi}\ e^{j\ \Gamma_z z}\ e^{jwt}$$

$H_n^{(1)}$, $H_n^{(2)}$ -- Hankel functions

$$k^2 + \Gamma_z^{\ 2} = \Gamma_o^{\ 2}; \quad \Gamma_o = 2\pi / \lambda_o$$

Circularly symmetrical waves: $n = 0$

$$H_o^{(1)} \quad \text{and} \quad H_o^{(2)}$$

These functions describe traveling waves caused by line sources. For large values of kr, the waves become plane waves

$$H_o^{(2)}\ (kr) \quad \rightarrow \sqrt{\frac{2j}{\pi kr}}\ \ e^{-jkr}$$

c. Spherical coordinates

Solution of the wave equation

$$\phi\ 1,2 = \begin{bmatrix} h_n^{(1)}\ (kr) \\ \\ h_n^{(2)}\ (kr) \end{bmatrix} P_n^{\ m}\ (\cos\theta)\ e^{jm\phi}\ e^{jwt}$$

$h_n^{(1)}$, $h_n^{(2)}$ --- spherical Hankel functions

P_n^m associated legendre functions of the first kind. These wave functions describe waves originating from a point source. For large values of kr, the waves have the form of plane waves:

$$h_o^{(2)} (kr) \approx - \frac{e^{-jkr}}{jkr}$$

III. Relationships Between Wave Functions in Different Coordinates (Review).

a. Mathematical approach by expressing a specific wave function by another one by mathematical manipulation

1. Plane wave functions expressed by cylindrical waves

$$e^{jkz} = e^{-jk\rho \cos\phi} = \sum_{n=-\infty}^{\infty} a_n J_n (k\rho) e^{jn\phi}$$

$$J_n = (H_n^{(1)} + H_n^{(2)}) /2 \text{ -- standing cylindrical waves}$$

2. Cylindrical waves expressed by plane waves

Multiply the above equation by $e^{-jm\phi}$ and integrate from 0 to 2π.

$$2\pi\, a_n J_n (kr) = \int_o^{2\pi} e^{-jkr \cos\phi} e^{-jm\phi} d\phi \quad ; \quad a_n = j^{-n}$$

3. Spherical waves expressed by cylindrical waves

$$\frac{e^{-jkr}}{r} = \frac{1}{2j} \int_{-\infty}^{\infty} H_o^{(2)} (r\sqrt{k^2 - u^2})\, e^{juz} du$$

4. Plane waves expressed by spherical waves

$$e^{-jkz} = e^{-jkr} \cos\phi = \sum_{n=0}^{\infty} j^n (2n+1)\, j_n (-kr) \quad P_n (\cos\theta)$$

$$j_n (\rho) = \sqrt{\pi/2\rho} \quad J_{n+1/2} (\rho) \text{ -------spherical Bessel function.}$$

b. Physics Approach

Example: Spherical waves expressed by cylindrical wave functions. Assume a constant harmonically varying current on an infinite wire along the z-axis of a cylindrical coordinate system. The radiation field can be described by

$$A_1 = C_1 H_o^{(2)} (\Gamma_o \rho).$$

If waves travel along the wire, we write $e^{j\Gamma_z z}$ and

$$A_2 = C_2 H_o^{(2)} (k\rho); \; k^2 + \Gamma_z^2 = \Gamma_o^2$$

A point source of a current can be expressed by a delta function

$$I\,(z) \;=\; \delta(z) \;=\; \int_{-\infty}^{\infty} e^{-j\Gamma_z z} \, d\Gamma_z$$

The radiation by the point source is hence given by

$$A_3 = C_2 \int_{-\infty}^{\infty} H_o^{(2)} \left(\rho\sqrt{\Gamma_o^2 - \Gamma_z^2}\right) e^{-\Gamma_z z} \, dz$$

The far-field of a point source is also

$$A_3 \;=\; \frac{e^{-j\Gamma_o r}}{r}$$

This leads to the same relationship as found by pure mathematical considerations.

IV. Wave Propagation in Nonuniform Media (Reference [3]).

(Figure 3)

$$\nabla \times \bar{E} = -jw\,\mu_o \bar{H},$$

$$\nabla \times \bar{H} = jw\,\epsilon_o \epsilon_r (\bar{r}\,') \bar{E} + \bar{J}\,(\bar{r}\,'),$$

$$\nabla \cdot \bar{B} = 0;\ \mu_r = 1;\ \ \nabla \cdot \bar{H} = 0,\ \bar{B} = \mu_o \bar{H},$$

$$\nabla \cdot \bar{D} = 0,\ \nabla . [\epsilon_o \epsilon_r (\bar{r}\,') \quad \bar{E}] = 0,$$

$$\epsilon_o \epsilon_r (\bar{r}\,') \quad \cdot E + \quad [\,\epsilon_o \epsilon_r (\bar{r}\,')] \cdot \bar{E} = 0$$

The following identities are introduced:

$$\nabla \times \nabla \times \bar{E} = \nabla (\nabla \cdot \bar{E}) - \nabla^2 \bar{E},$$

$$\nabla \times \nabla \times \bar{E} = -\nabla \left[\frac{\nabla \epsilon_r (\bar{r}\,') \cdot \bar{E}}{\epsilon_r (\bar{r}\,')} \right] - \nabla^2 \bar{E},$$

$$\nabla \times \nabla \times \bar{H} = -\nabla^2 \bar{H}$$

It follows:

$$\nabla^2 \bar{E} + \beta_o^2 \bar{E} = jw\,\mu_o \bar{J} - \nabla [\nabla \epsilon_r (\bar{r}\,') \cdot \bar{E}\,] - \beta_o^2\ \Delta \epsilon_r (\bar{r}\,')\ \bar{E},$$

$$\nabla^2 \bar{H} + \beta_o^2 \bar{H} = -jw\,\epsilon_o \quad [\nabla \epsilon_r (\bar{r}\,') \times \bar{E}\,] - \beta_o^2 \Delta \epsilon_r (\bar{r}\,') \bar{H} - \nabla \times \bar{J},$$

Where $\epsilon_r (\bar{r}\,') = 1 + \Delta \epsilon_r (\bar{r}\,'),\ \beta_o^2 = w^2\ \epsilon_o \mu_o = \frac{w^2}{c^2}$, $c = 3 \times 10^8$ m/sec.

The nonhomogeneous wave equation is solved by a Hertz vector

$$\bar{E} = \nabla(\nabla\cdot\bar{\Pi}) - \epsilon_o\mu_o\frac{\partial^2\bar{\Pi}}{\partial t^2}, \quad \bar{\Pi}(\bar{r}, t) = \frac{1}{4\pi\epsilon_o}\int_o^t dt\int_v \frac{\bar{J}(\bar{r}', t - \frac{\Delta r}{c})}{\Delta r}\, dv,$$

$$\text{(Where } \bar{A} = \epsilon_o\mu_o\frac{\partial\bar{\Pi}}{\partial t^2}\text{)}, \quad \bar{E} = \nabla(\nabla.\bar{\Pi}) + \beta_o^2\bar{\Pi}, \quad \Delta r = |\bar{r} - \bar{r}'|,$$

$$\bar{\Pi}(\bar{r}) = (4\pi jw\epsilon_o)^{-1}\int_V \frac{\bar{J}(\bar{r}', t - \Delta r/c)}{|r - r'|}dv = (4\pi jw\epsilon_o)^{-1}\int_v \frac{\bar{J}(\bar{r}')e^{-j\beta_o|r - r'|}}{|r - r'|}dv$$

Substitution yields

$$\bar{E} = (4\pi jw\epsilon_o)^{-1}\int_V\left[\nabla(\nabla\cdot\bar{J}_{tot}\frac{e^{-j\beta_o|\bar{r} - \bar{r}'|}}{|\bar{r} - \bar{r}'|}) + \beta_o^2\bar{J}_{tot}\frac{e^{-j\beta_o|\bar{r} - \bar{r}'|}}{|\bar{r} - \bar{r}'|}\right]dv,$$

Where $\bar{J}_{tot} = jw\mu_o\bar{J} - \nabla[\nabla\epsilon_r(\bar{r}')\cdot\bar{E}] - \beta_o\Delta\epsilon_r(\bar{r}')\bar{E}$.

In simplified form, $\bar{E}$ can be written

$$\bar{E}(\bar{r}) = \int_V \bar{\bar{\Gamma}}(\bar{r}, \bar{r}')\bar{J}_{tot}\,dv, \quad \bar{\bar{\Gamma}} = (4\pi jw\epsilon_o)^{-1}(\beta^2 + \nabla\nabla)\frac{e^{-j\beta_o|\bar{r} - \bar{r}'|}}{|\bar{r} - \bar{r}'|}.$$

The integral equation can be solved by writing

$$\bar{E}(\bar{r}) = \bar{E}_o(\bar{r}) + \sum_{n=1}^{\infty}\Delta\bar{E}_n(\bar{r}).$$

and deriving a recursion formula.

Example: Doppler Shift in Nonuniform Media

The field intensity is

$$E(\bar{r}, t) = A(\bar{r})\ e^{i\,\phi(\bar{r})jwt},$$

Where

$$\log A(\bar{r}) = \mathrm{Re}\left[\int_{o}^{S} \frac{dE(\bar{r})}{ds'} \Big/ E(\bar{r})\, ds'\right],$$

$$\phi(\bar{r}) = \mathrm{Im}\left[\int_{o}^{S} \frac{dE(\bar{r})}{ds'} \Big/ E(\bar{r})\, ds'\right].$$

The Doppler Shift is

$$\Delta w = \frac{d\,\phi(\bar{r})}{dt} = \frac{d\phi(\bar{r})}{d\bar{r}}\,\frac{d\bar{r}}{dt} = \nabla\phi \cdot \bar{v}.$$

If s has the direction of the velocity of the source $\bar{v}$ we find

$$\Delta w = \mathrm{Im}\ \frac{dE(\bar{r})}{d(\bar{v}t)} \Big/ E(\bar{r}) .$$

The equation permits representation of the Doppler Shift in closed form.

V. Wave Propagation in Plasma (Reference [4] and [5]).

1. Plasma Properties

The properties of an isotropic transmission medium can be described by medium constants which interrelate the field quantities.

$$\overline{D} = \epsilon \overline{E} \quad , \quad \overline{B} = \mu \overline{H} \ ,$$

where

$$\epsilon = \epsilon_o \epsilon_r = \epsilon_o (\epsilon_r' - j \ \epsilon_r'') \ ,$$

$$\mu = \mu_o \mu_r = \mu_o (\mu_r' - j \mu_r'') \ .$$

We use Maxwell's equation

$$\nabla \times \overline{H} = J_{tot} = \frac{\partial \overline{D}}{\partial t} + \overline{J} = (jwE + \sigma) \overline{E} = jw \epsilon_o \epsilon_r \overline{E} \ ,$$

where $\epsilon_r = \epsilon_r' - j \frac{\sigma}{w \epsilon_o} = \epsilon_r' - j \ \epsilon_r'' = \frac{\sigma}{w \epsilon_o}$.

For anisotropic media, ϵ_r becomes a tensor,

$$D_x = \epsilon_{11} E_x + \epsilon_{12} E_y + \epsilon_{13} E_2 \ ,$$

$$D_y = \epsilon_{21} E_x + \epsilon_{22} E_y + \epsilon_{23} E_z \ ,$$

$$D_z = \epsilon_{31} E_x + \epsilon_{32} E_y + \epsilon_{33} E_z \ ,$$

so that

$$| \epsilon_r | = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \ .$$

Expressing the tensor permittivity by a conductivity yields

$$\bar{\bar{\epsilon}}_r = \bar{\bar{1}} + \frac{\bar{\bar{\sigma}}}{jw\,\epsilon_o} ,$$

Where $\bar{\bar{1}}$ is the unit diadic.

In the case of a plasma

$$\bar{J}_{tot} = jw\,\epsilon_o\,\bar{E} + \sum_{n=1}^{N} q_n\,\bar{v}_n \,.$$

The second term is due to the motion of the N particles, with the individual charges q_n and velocities $\bar{v}_n$. The motion is caused by the internal dynamics of the particles and by external forces

$$\bar{F} = q\,(\bar{E} + \bar{v} \times \bar{B}).$$

To take into account the interaction with each particle separately is impossible and simplified models have to be applied. Statistical methods and averaged quantities are commonly used.

Simplifying assumptions: Interaction with electrons only, Maxwell-Boltzmann velocity distribution, collision frequency is constant, quasi-static field interaction, static magnetic field in z-direction of a rectangular coordinate system. With these assumptions, $\bar{\bar{\epsilon}}_r$ is given by

$$|\epsilon_r| = \begin{bmatrix} \epsilon_r^1 & -j\epsilon_r^2 & 0 \\ j\,\epsilon_r^2 & \epsilon_r^1 & 0 \\ 0 & 0 & \epsilon_r \end{bmatrix}$$

Further relationships:

$$\epsilon_r^1 = 1 - \left(\frac{wp}{w}\right)^2 \frac{(jw + \upsilon)\, jw}{(jw + \upsilon)^2 + w_q^2} = 1 - p^2 \frac{1 - jq}{(1 - jq)^2 - 0^2} \; ,$$

$$i \; \epsilon_r^2 = -\left(\frac{wp}{w}\right)^2 \frac{jww_q}{(jw + \upsilon)^2 + w_q^2} .= - p^2 \frac{0}{(1 - jq)^2 - 0^2} \; ,$$

$$\epsilon_r = 1 - \left(\frac{wp}{w}\right)^2 \frac{jw}{jw + \upsilon} = 1 - p^2 \frac{1}{1 - jq} \; ,$$

plasma frequency $w_p = (Neq_e^2 / \epsilon_o m_o)^{1/2} \approx 56.5\, N_e^{1/2}$,

collision frequency $\upsilon = N_i \; \langle v_e \rangle \; Q_{ci}$

gyro frequency $w_q = \frac{qe}{m_e} \; B_z$

These equations are obtained by derivation from Boltzmann's equation and represent a simplified but useful model for the plasma properties.

2. Wave Propagation:

$$\nabla \times \bar{E} = - jw \mu_o \bar{H}, \nabla \times \bar{H} = jw \epsilon \bar{E} = jw \epsilon_o \bar{\bar{\epsilon}}_r \cdot \bar{E} \; ,$$

$$\nabla \times \nabla \times \bar{E} - \beta_o^2 \; \bar{\bar{\epsilon}}_r \cdot \bar{E} = 0.$$

Assume a solution of the vector wave equation in rectangular coordinates in the yz-plane with a magnetic field in the z-direction (B_z) (see Fig. 4)

$$\bar{E} = \bar{E}_o \ \exp\left[j(wt - \Gamma_y y - \Gamma_z z)\right] , \ \Gamma_{y,z} = \beta_{y,z} - j\,\alpha_{y,z} , \ \Gamma^2 = \Gamma_y^{\ 2} + \Gamma_z^{\ 2} .$$

We obtain

$$(\Gamma^2 - \beta_o^{\ 2} \ \epsilon_r^{\ 1}) \ E_x - j\,\beta_o^{\ 2} \ \epsilon_r^{\ 2} \ E_y \ + \ 0 \ = \ 0,$$

$$j\beta_o^{\ 2} \ \epsilon_r^{\ 2} \ E_x \ + \ (\Gamma_z - \beta_o^{\ 2} \ \epsilon_r') \ E_y \ - \ \Gamma_y \ \Gamma_z \ E_z \ = \ 0,$$

$$0 \ + \ \Gamma_y \Gamma_z \ E_y \ + \ (\Gamma_y^{\ 2} - \beta_o^{\ 2} \ \epsilon_r) \ E_z \ = \ 0.$$

Setting the determinant zero and introducing

$$\Gamma_z = \Gamma \cos\theta , \quad \Gamma_y = \Gamma \sin\theta$$

yields the dispersion relation

$$A \ \frac{\Gamma}{\beta_o}^{\ 4} \ - \ B \ \frac{\Gamma}{\beta_o}^{\ 2} \ + C = 0,$$

where

$$A = \epsilon_r \cos^2\theta \ + \ \epsilon_r' \ \sin^2\theta ,$$

$$B = \epsilon_r' \ \epsilon_r \ (1 + \cos^2\theta) \ + \ \left[(\epsilon_r^{\ 1})^2 \ - \ (\epsilon_r^{\ 2})^{\ 2}\right] \ \sin^2\theta,$$

$$C = \left[(\epsilon_r^{\ 1})^2 \ - \ (\epsilon_r^{\ 2})^2 \right] \ \epsilon_{r\varnothing} \ .$$

The solution is $\left(\frac{\Gamma}{\beta_o}\right)^2 = \dfrac{B \pm \sqrt{B^2 - 4AC}}{2A}$.

Index of refraction $n = \frac{\Gamma}{\beta_o}$, $\beta_o = 2\pi / \lambda_o$.

3. Typical cases of wave propagation.

a. No magnetic field:

$$w_q = 0, \epsilon_r^1 = \epsilon_r , \epsilon_r^2 = 0,$$

$$A = \epsilon_r , \quad B = (\epsilon_r)^2 , \quad C = (\epsilon_r)^3 .$$

$$\epsilon_r = 1 - \frac{p^2}{1 - jq} ,$$

ϵ_r --- equivalent relative permittivity of the plasma.

b. Propagation along a magnetic field:

$$\theta = 0 ,$$

$$A = \epsilon_r , \quad B = 2 \epsilon_r^1 \epsilon_r , \quad C = \left[(\epsilon_r^1)^2 - (\epsilon_r^2)^2 \right] \epsilon_r ,$$

$$\frac{\Gamma}{\beta_o}^2 = (\epsilon_r^1 \pm \epsilon_r^2) .$$

The typical wave modes which are circularly polarized, have different wave velocities and propagation constants Γ .

$$(\Gamma_{right})^2 = \beta_o^2 \epsilon_R = \beta_o^2 (\epsilon_r^1 - \epsilon_r^2),$$

$$(\Gamma_{left})^2 = \beta_o^2 \epsilon_L = \beta_o^2 (\epsilon_r^1 + \epsilon_r^2),$$

$$\epsilon_R \epsilon_L = (\epsilon_r^1)^2 - (\epsilon_r^2)^2 ,$$

$$\epsilon_{R\ (L)} = 1 - \frac{p^2}{(1 \pm o) - jq} .$$

C. Propagation across magnetic fields:

$$\theta = \pi/2,$$

$$A = \epsilon_r^1 , \quad B = \epsilon_r^1 \epsilon_r + \left[(\epsilon_r^1)^2 - (\epsilon_r^2)^2 \right] ,$$

$$C = \left[(\epsilon_r^1)^2 - (\epsilon_r^2)^2 \right] \epsilon_r ,$$

$$\Gamma_\perp^2 = \beta_o^2 \epsilon_r ,$$

$$\Gamma_{||}^2 = \beta_o^2 \epsilon_r^1 \left[1 - \left(\frac{\epsilon_r^2}{\epsilon_r^1} \right)^2 \right] .$$

References

1. J. A. Stratton, "Electromagnetic Theory," McGraw-Hill Book Co., Inc., New York, 1941.

2. R. F. Harrington, "Time-Harmonic Electromagnetic Fields," McGraw-Hill Book Co., Inc., New York, 1961.

3. F. J. Tischer, "Doppler Phenomena in Space Communications," IRE Transactions Communication Systems, vol. CS-7, no. 1, 25-30, May 1959.

4. F. J. Tischer, "Wave Propagation Through Ionized Gas in Space Communications," IAS Report No. 59-34, Institute of the Aeronautical Sciences.

5. F. J. Tischer, "Wave Propagation Through a Hot Plasma," in Advances in the Astronautical Sciences, vol. 8, Plenum Press, New York, 1963.

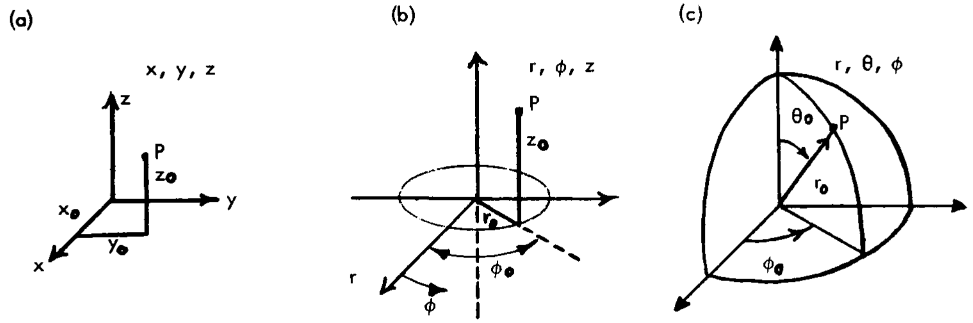
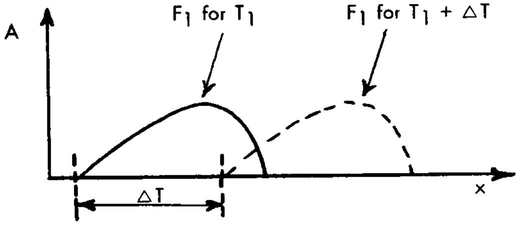


Fig. 1 Coordinate systems



$$F_1\left(t + \Delta T - \frac{z + \Delta z}{v}\right) = F_1\left(t - \frac{z}{v}\right) \text{ if } \Delta T = \frac{\Delta z}{v}$$

Fig. 2 Waves traveling in the positive x-direction

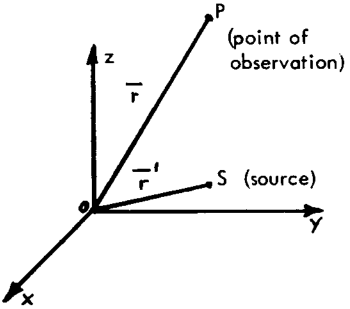


Fig. 3

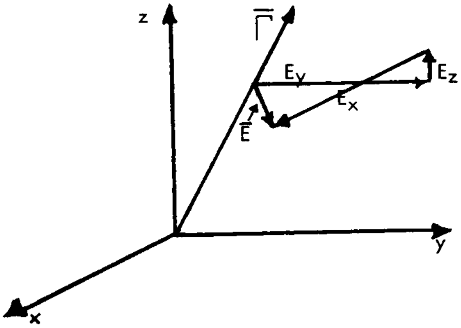


Fig. 4

SPECIALIZED TOPICS I

Radio Astronomy and Its Relationship to Space Communications

(Abstract)

by William C. Erickson

From an observational standpoint, radio waves are a powerful tool for exploring the universe for two reasons; radio receiving systems are far more sensitive than optical sensors so they can detect fluxes of radiation many orders of magnitude lower than those detectable optically, and radio waves propagate unhindered through clouds of interstellar material which extinguish optical radiation. From a theoretical standpoint, radio observations allow us to examine a whole range of energetic processes which occur in the universe and which are not detectable by other means. Thus, radio observations yield data concerning a much larger sample of the universe than that observable by other means, and yield data concerning energetic processes which may be the key to many of the most intriguing problems in astrophysics.

In relation to space communications systems, galactic background radiation is the primary component of total system noise at low frequencies. At L-band and above, this is not the case, but with high gain antennas the noise intensity received in certain directions from discrete radio sources can be important. Solar emission can cause serious disturbances under any conditions. At long wavelengths the scattering of radio waves by interplanetary electron clouds should be considered.

2425502 Weather Bureau, Washington, D.

t', Earth Environment Effects

by S. Fred Singer In..., p 52-63 ref

(see...,)

Introduction

The magnetosphere of the earth covers the region where ionized gas (plasma) plays an important part in the dynamics of the atmosphere and where the geomagnetic field therefore occupies an important role. The magnetosphere begins (by convention) at the maximum of the F layer of the ionosphere, at about 250 km, and extends out to perhaps 10 to 15 earth radii, at which point the interplanetary plasma or solar wind limits the earth's magnetosphere. (For reference, the earth's radius is 6,370 km (4,000 miles).

We outline here the properties of the magnetosphere as they are known and of phenomena occurring in the magnetosphere; but it should be stressed that all throughout the dynamics are very important, but hardly known. Time variations occur everywhere in the magnetosphere, e.g., in the density and composition at altitudes as low down as 500 km. For the most part, however, we will be talking about "mean" properties, with only slight reference to these time variations.

Thermosphere

Beginning at about 80 km the O_2 molecule is rapidly dissociated by solar ultraviolet in the Schuman Runge bands; by about 120 km to 150 km, the process is nearly complete. Nitrogen, because of its higher dissociation energy remains mainly in molecular form. Thus, at 250 km the major constituents are atomic oxygen and molecular nitrogen. Most of the ions are in the form of O^+. Turbulent mixing of the atmosphere is believed to stop somewhere between 110 km and 150 km, so that gravitational diffusion predominates above these altitudes. As a consequence, the heavy gases, including also N_2, will become increasingly less important with increasing altitude, while light gases, such as He and H play a more important role. At 250 km these are still minor constituents, but at much greater altitudes, as we shall see, they eventually predominate.

Densities and scale heights in the region 250 to 600 km are now well-known from satellite drag measurements, as well as from more direct measurements. Coupled with a knowledge of atmospheric composition, which gives mean molecular weight, it is possible to deduce the temperature. This region, sometimes called the thermosphere has a reasonably constant temperature, since there is little gas to absorb further energy from solar ultraviolet radiation and the heat loss, mainly due to the coronal transition of atomic oxygen at 62 microns, becomes rather small and is counteracted by conduction from below. Among the most interesting processes going on in the thermosphere is the upward diffusion of helium and hydrogen atoms which percolate rapidly to the 500 to 600 km level.

Neutral Exosphere

At this level, we generally place the base of the exosphere, a region from which escape of neutral atoms becomes possible. There is, of course, no sharp boundary, but by convention, the base is taken to be the level at which the "mean free path" in horizontal direction equals the local scale height. Other definitions are possible or even preferable; for example, that one-half of the atoms issuing in an upward direction with more than escape velocity, will escape without making any further collisions. Hydrogen atoms escape most rapidly since their RMS speed at a temperature of about 1500^{o} is around 6 km per second, so that an appreciable Maxwellian tail exists having more than the escape velocity. The escape is so rapid, in fact, that it is limited by diffusion from below. This fact has only recently been realized. It has an important bearing on the diurnal variation in the density of hydrogen in the exosphere, as well as on density variations during the solar cycle. But, in particular, it affects the density at all times for hydrogen at distances of more than 10 earth radii, since that population is supplied from the extreme end of the Maxwellian tail which is replenished only with great difficulty. Helium also escapes but at a lesser rate, the helium 4 isotope less rapidly than the helium 3 isotope. The bookkeeping operation which establishes the relative rates of escape and rates of production for the two isotopes has not yet been properly done and is still a challenging field of research.

Atomic oxygen is the predominant constituent in the lower exosphere; very few atoms escape. The remainder all describe simple Keplerian trajectories in the earth's gravitational field.

Aside from ballistic trajectories (portions of ellipses) and escape orbits which are hyperbolic, there also exists the possibility of atoms in elliptic satellite orbits, placed there by a variety of means. There exists some divergence of opinion about the actual number of, e.g. hydrogen atoms in bound orbits at a distance of several earth radii. This is an important point since the bound orbits, if they are present, would increase the density of hydrogen considerably. In various observations on the scattering of Lyman alpha radiation the question of hydrogen densities at large distances enters in an important way. It also enters importantly in determining the lifetime of the magnetic storm ring current (See below).

For neutral atoms in the exosphere one cannot really define a temperature since the Maxwellian distribution is modified and chopped up; however, in practice this raises no problem since the density is the important parameter. An "effective" temperature however is determined by the temperature of the thermosphere. The latter is known to vary and as a consequence the densities, compositional ratios, etc. in the exosphere will also vary, but this subject has not been fully developed either theoretically or experimentally.

Magnetospheric Plasma

Quite a separate topic is the status of the ionized gas in the region of the exosphere; i.e., the magnetospheric plasma, properly speaking. Because of its very large collision cross section, essentially the Coulomb cross section, ions do not form an exosphere until their density drops below something like ˛200 per cm^3. The problems of an iono-exosphere have only been treated theoretically, but show that the ions cannot escape because of the earth's magnetic field. Instead, there exists two types of orbits, those that go over the top through the equatorial plane and those that are brought back by the gravitational field before they reach the equatorial plane. In the polar regions the problem of escape is less well-defined; it may be possible for ions to move out along lines of force into interplanetary space, but the topology of these lines is not well enough known. (Conversely, interplanetary plasma can move along these lines, of course, down into the earth's thermosphere.)

Throughout most of the magnetosphere the ions are in thermal equilibrium because of frequent collisions and form essentially a barosphere that obeys the hydrostatis equation (unlike the neutral atoms). With frequent collisions an ion can drift across lines of force and is no longer bound to a particular line, Therefore, the influence of the magnetic field is not important. On the other hand, it has been shown that the differential diffusion resulting from the very light mass of the negative electrons produces an electrostatic field, radially oriented, which acts to reduce the gravitational force on the ions and therefore "pulls them up." As a result, the concentration of the light ions, such as He^+ and H^+, will increase with altitude in the presence of the predominant ion, O^+. This process stops when the O^+ density has fallen to a sufficiently low value where He^+ becomes predominant, and it will then fall off according to a quasi-barometric law, the exact nature of which has not yet been delineated, either theoretically or experimentally. The interesting possibility has been raised of the existence of doubly and even triply-ionized ions, such as He^{++} and O^{++} and O^{+++}. These highly charged ions will be "pulled up" by the electrostatic field so that recombination, and, therefore, their rate of disappearance, becomes quite small, thus increasing their equilibrium concentration. Their existence has not yet been demonstrated experimentally. The temperature of the thermosphere is determined mainly by the absorption of solar ultraviolet. This is shown quite clearly by the large diurnal variation in temperature. While other heating sources may occasionally be important, such as hydromagnetic waves, and corpuscular radiation in the auroral zone, it seems clear that the absorption of solar ultraviolet is of overriding importance. The absorption of a UV photon creates a photo-electron having considerable energy. Many of these move in an upward direction and because of their high velocity (and small Coulomb cross section), they are effectively trapped by the magnetic field and follow the line of force to the other hemisphere. As a result, there must exist a corona of hot electrons in the earth's magnetosphere. These electrons may provide the seeds for the high energy trapped electrons, essentially by allowing themselves to be accelerated by various types of hydromagnetic waves in the magnetosphere. Then again, the hot electrons are important also since they determine the electric charge of bodies moving in the magnetosphere.

Various authors have pointed to the possibility of convection in the magnetosphere. Some picture it in terms of instabilities, but there is no agreement on this point. Others have discussed an ordered convection based on external electric fields due to the interaction of the earth's magnetic field with the solar wind. One model of magnetic storms is based on such a convection theory.

Interplanetary Dust Particles

The earth's gravitational field causes the deflection of orbits of small interplanetary dust particles, some of which are even accreted by the earth. A number of theoretical investigations have been made which, surprisingly, lead to different results. The writer believes that the gravitational field will increase the dust concentration by a small factor, causing a maximum at an altitude of about 2,000 km above sea level.

On the other hand, the measured quantity, the flux, of dust particles will show a very pronounced increase, something like two or three orders of magnitude, with a peak at 2,000 km, and thus form a dust shell. The experimental results so far seem to bear out this point of view. A further theoretical prediction is a morning to evening asymmetry of the flux of very small dust particles in the vicinity of the earth. This would be produced through the effects of solar radiation pressure. The question of whether dust particles exist in satellite orbits about the earth has been raised but has not been settled either theoretically or experimentally.

Geomagnetic Field

The earth's magnetic field is quite complicated at sea level since the local anomalies are still quite pronounced. At larger distances from the earth, the approximation of an eccentrically placed dipole is adequate for many purposes. However, for refined calculations, for example, on the lifetime of radiation belt particles, it is often necessary to use the actual field. Tables have been prepared giving the positions of field lines and evaluating the magnitude of the field along the field line, giving the longitudinal invariants and other information necessary for the purpose.

Trapped Radiation

High-energy particles trapped in the earth's magnetic field constitute one of the interesting phenomena in the magnetosphere. For purposes of discussion, we will take them in four groups.

1. Protons between 10 to 700 Mev. These protons are located mainly in the lower part of the magnetosphere along lines of force which extend up to a latitude of about 50^{o}. The most successful explanation of their origin is in terms of the decay protons from albedo neutrons produced by cosmic rays which enter the earth's atmosphere. This neutron albedo theory seems to account quite well for the spatial distribution of protons, for the observed energy spectrum, and even for some of the observed time variations. The latter seem to be due to the time variations in the influx of cosmic rays, particularly the large bursts associated with solar flares. These enter in the polar regions, produce neutron albedo by disintegration of atmospheric nuclei; the neutron albedo, in turn, enhances the trapped proton intensities. The neutron albedo theory predicts very well the altitude dependence of trapped protons and shows that it is roughly inversely proportional to the atmospheric density up to altitudes of a few thousand kilometers. In fact, it has become possible to use the measurements on trapped protons to deduce atmospheric scale heights and densities. The neutron albedo theory predicts rather long lifetimes for high energy protons, of the order of several hundred years, at altitudes of 2,000 - 3,000 km. At much higher altitudes, the trapping properties of the magnetic field apparently are weakened and the lifetime of the high energy protons is considerably reduced,

leading to a reduced intensity. As a result, the trapped protons show a maximum at an altitude of about one-half earth radius.

2. Protons with energy below 10 Mev originate at least partly from the decay of albedo neutrons, but there may also be another source, mainly, low energy protons from interplanetary space which, in turn, may become accelerated in the earth's magnetic field. Not too much is known yet about the detailed distribution of very low energy protons in the magnetosphere, with energies below 100 Kev.

3. High energy electrons, i.e., those having an energy greater than 300 Kev, are only partly accounted for by cosmic ray albedo neutrons, which produce electrons in beta decay. Since the flux of lower energy electrons is so great, it suggests that another source is present, possibly the acceleration of the high energy tail of magnetospheric plasma. Large time variations in intensity are observed, correlated with magnetic storms, and suggesting a relationship which, however, has not been unraveled. It is clear that complicated acceleration processes are present, and it is most likely that the ultimate energy for these comes from the Sun, probably in the form of shock waves transmitted through the interplanetary plasma into the earth's magnetosphere.

4. The low energy electrons are largely responsible for the observed aurora. The energy flux into the auroral zones is much too large to be accounted for in terms of a cosmic ray origin for the electrons. This further supports the idea for another mechanism which furnishes trapped electrons.

Magnetic Storms

Magnetic storms are a complex phenomenon whose cause is believed to lie in the magnetosphere. It sometimes starts with a "sudden commencement," which may in some cases even be preceded by a "reverse sudden commencement." There are two schools of thought as to the location of the currents causing this magnetic disturbance, and it would be important to know whether the currents are indeed located in the ionosphere as some workers believe. The "main phase" of the magnetic storm which follows the sudden commencement and lasts for a day or so is now generally believed to be due to a ring current made up of trapped particles. The nature of the particles is not certain. They could either be low energy protons which are then removed by charge exchange with neutral hydrogen; on the other hand, they could also be low energy electrons, which are removed by scattering. Various experimental tests have been proposed to distinguish between these two possibilities, but they have not as yet been carried out.

Finally, we come to discuss the outer limits of the earth's magnetosphere. Its limits have been determined both by magnetic measurements and plasma measurements, as well as by measurements of the trapped particles. They suggest a teardrop shape with a flattening in the direction of the Sun due to the pressure of the solar wind. It is believed that a fixed shock wave, a bow wave, exists at some distance in front of the boundary and that between the stationary shock front (which must be collision-less) and the actual cavity boundary, the solar plasma, flowing through the shock front, may thermalize and then flow along with the boundary of the geomagnetic cavity. The estimates from, roughly, seven satellite and space-probe flights so far would indicate that the cavity on the sunlit side extends to about 10 earth radii and that the shock front is located at about 15 earth radii. It is clear, however, that a proper exploration of this difficult boundary region can be obtained only by repeated meausrements. Time variations certainly do exist and should be related to magnetic storm effects observed within the magnetosphere and on the earth.

Structure of the Ionosphere

(Abstract)

by R. E. Bourdeau

Results of rocket and satellite experiments designed to study the characteristics of charged particles in the ionosphere were discussed and compared with ground-based ionospheric observations. Models were presented of the ionosphere to altitudes of 2,000 km which includes the D, E, and F regions as well as the heliosphere and the base of the protonosphere.

SPECIALIZED TOPICS II

Coherence Properties of Light

(Abstract)

by E. Wolf

In the first part of this lecture, an introductory account was given of the basic concepts relating to the description and to the analysis of optical coherence effects.

In the second part, a review was given of some of the more recent developments in this area of physics. It dealt mainly with higher order coherence effects, intensity interferometry, photon coincidence experiments, transient interference from independent sources, statistical properties of laser light and quantum theory of coherence.

California U., San Diego

t: Quantum Electronics and Communication

by

James P. Gordon*

In... p 65-72 refs

(See...)

With the rise of quantum electronics in the last ten years or so, and in particular with the creation of ultra low-noise amplifiers (masers, parametric amplifiers), in the microwave frequency range, and of coherent oscillators and amplifiers in the near infra-red and optical regions of the spectrum, it has become important to extend the basic principles of communication theory to include the no-longer-negligible effects of quantization. Quantization of the radiation field provides fundamental limits on the rate at which information can be transferred from a sender to a receiver with a given amount of power, and moreover provides limitations on the range of frequencies which are useful for communications.

Let us examine these fundamental limitations. To do this, let us assume that there are no "practical" limitations on what we can do: that we have "ideal" transmitting and receiving equipment at all frequencies. In this context the concept of entropy is very important. C. E. Shannon, N. Weiner, and others recognized, about 1948, the close connection between information and entropy. Entropy is a measure of the randomness of the disposition of the parts that make up a physical system. Information (or more precisely, information capacity), is the same measure of an apparent randomness which is, however, controlled by a transmitter, and interpretable by a receiver. The greatest information which can be incorporated in a physical system is in fact its total entropy less that part of the entropy which is either uncontrollable or uninterpretable. The uncontrollable randomness of the physical system is what is usually thought of as noise. In addition, however, quantum-mechanical limitations on measuring processes by receivers introduce further restrictions on information which can be important but are often not easy to assess.

Let us now apply these ideas to a basic communication system which consists of a transmitter, a transmission medium, and a receiver. For simplicity, we will assume that a single transverse mode of the radiation field is utilized. That is, the field distribution and polarization over any plane perpendicular to the direction of propagation is to be considered invariant. Then to describe the received field we need only examine its temporal variation. The field which is received in some reasonably long time T may be analyzed into the complete set of Fourier components which are pericdic in T; i.e., for which

$$\nu_n T = n ,$$

n being an integer. Hence the total field strength may be written

$$E(t) = \mathrm{Re} \sum_{n=1}^{\infty} a_n e^{-i2\pi\nu_n t} = \sum_{n=1}^{\infty} |a_n| \cos(2\pi\nu_n t + \varphi_n)$$

where Re indicates that the real part of the expression is to be taken. The values of the complex numbers a_n thus determine the field completely. In a small frequency range $\delta\nu$, then are clearly

$$\delta n = T\delta\nu$$

Fourier components, or "modes."

Suppose now that the received field has an average spectral power density, S_ν near frequency ν, so that the average received power in the small frequency band $\delta\nu$ is

$$S_\nu \delta\nu$$

The average energy per mode is then simply S_ν. Quantum mechanically, the entropy of the field in the range $\delta\nu$ is maximum when the phases φ_n are completely random and when the probability of finding a certain amount of energy in a mode is an exponentially decreasing function of that energy. In addition, the mode energy must be quantized in units of $h\nu$, where h is Planck's constant. The mathematical expression of this is

$$\text{prob}(n) = \frac{1}{1 + \bar{n}} \left(1 + \frac{1}{\bar{n}} \right)^{-n}$$

where $\bar{n}$ is the average of n and hence is equal to $S_\nu/h\nu$. The mathematical expression for the quantum mechanical entropy per mode is:

$$\begin{aligned} H &= - \sum_{n=0}^{\infty} \text{prob}(n) \log \left[\text{prob}(n) \right] \\ &= \log (1 + \bar{n}) + \bar{n} \log (1 + \frac{1}{\bar{n}}) \\ &= \log \left(1 + \frac{S_\nu}{h\nu} \right) + \frac{S_\nu}{h\nu} \log \left(1 + \frac{h\nu}{S_\nu} \right) \end{aligned}$$

A useful quantity is the entropy rate δR for the frequency range $\delta\nu$, which is

$$\delta R = \frac{H\delta n}{T} = H\delta\nu = \delta\nu \log\left(1 + \frac{S_\nu}{h\nu}\right) + \frac{S_\nu \delta\nu}{h\nu} \log\left(1 + \frac{h\nu}{S_\nu}\right)$$

This expression is an upper limit to the rate at which information can be transmitted in the frequency range $\delta\nu$ when the average received power is $S_\nu \delta\nu$. The actual information is reduced from this limit by noise in the channel and by receiver limitations. When one adds to the transmitted field a Gaussian noise field with an average spectral power density N_ν, one finds, for the total entropy less the noise entropy the expression

$$\delta R = \delta\nu \log\left(1 + \frac{S_\nu}{N_\nu + h\nu}\right) + \frac{\left(S_\nu + N_\nu\right)\delta\nu}{h\nu} \log\left(1 + \frac{h\nu}{S_\nu + N_\nu}\right) - \frac{N_\nu \delta\nu}{h\nu} \log\left(1 + \frac{h\nu}{N_\nu}\right)$$

This expression then forms an upper limit to the information rate of communication channels when additive Gaussian noise is present. Note that the first term of this expression is dominant when either

$N_\nu \gg h\nu$ or $S_\nu \gg h\nu + N_\nu$; i.e. when $S_\nu + N_\nu \gg h\nu$.

It is interesting to compare this upper limit with what can be done with various possible receivers.

If the first element of the receiver is an ideal high gain linear amplifier (examples are masers (or lasers), parametric amplifiers) or an ideal high gain linear converter (heterodyne), there appears in the receiver a certain amount of noise, as a result of spontaneous emission in the amplifier or shot noise in the converter. For an input signal S_ν and noise N_ν, the output signal to noise ratio is

$$\frac{S_\nu}{N_\nu + h\nu}$$

which implies, by the classical information theory, an information capacity δC_{amp} of

$$\delta C_{amp} = \delta\nu \log\left(1 + \frac{S_\nu}{N_\nu + h\nu}\right)$$

As must be

$$\delta C_{amp} < \delta R$$

However, when $S_\nu + N_\nu >> h\nu$; that is, when the average energy per mode is much greater than one photon, communication systems involving good linear amplifiers or converters have capacities which approach closely to the theoretical limit.

Of possible importance to long distance space communications is the situation which occurs when $S_\nu + N_\nu$ is of order $h\nu$ or less. Let us look at the simple case $N_\nu << S_\nu << h\nu$. Here we would expect that the linear amplifier is not very efficient, since

$$\begin{aligned}\delta C_{amp} &\cong \delta\nu \log\left(1 + {}^{S}\nu/h\nu\right)\\ &\cong \frac{S_\nu \delta\nu}{h\nu} \log e\end{aligned}$$

while

$$\delta R \cong \frac{S_\nu \delta \nu}{h\nu}\left[\log e + \log\left(1 + \frac{h\nu}{S_\nu}\right)\right]$$

which can be considerably larger. In this case we have only a few photons at the receiver to go around among many modes; an information capacity approaching the entropy rate can be achieved by concentrating the transmitted energy into a few frequencies, chosen according to the message, and then, at the receiver, detecting which frequencies had been sent by means of a system of passive filters and energy detectors (i.e., quantum counters). The information rate for such a system approaches

$$\delta C_{\text{counter}} \longrightarrow \frac{S_\nu \delta \nu}{h\nu} \log\left(\frac{h\nu}{S_\nu}\right)$$

for very small values of $S_\nu/h\nu$.

The choice of which frequency band to use for best communication depends of course on the available transmitter power, the efficiency of producing this power and on the geometry of the system. Let us here restrict the discussion to the case $S_\nu + N_\nu >> h\nu$, so that the entropy rate and the information capacity assume the simpler forms

$$\delta R \gtrsim \delta C_{\text{amp}} = \delta \nu \log\left(1 + \frac{S_\nu}{N_\nu + h\nu}\right)$$

If N_ν is taken as black body radiation with an effective temperature T, then we have

$$N_\nu = \frac{h\nu}{\exp\left(\frac{h\nu}{kT}\right) - 1}$$

The total effective receiver noise is for this case

$$N_\nu + h\nu = \frac{h\nu}{1 - \exp\left(-\frac{h\nu}{kT}\right)}$$

This is a monatomically increasing function of frequency, starting at kT for the range $\nu \ll kT/h$, and becoming proportional to frequency for $\nu > kT/h$. Thus if S_ν were independent of frequency we would automatically want to use the lowest frequency which was compatible with the bandwidth requirements of the system. However, if communication takes in space, let us say, between two distant antennae, then under the assumption of constant sized transmitting and receiving antennae, and an available transmitter power independent of frequency, the received power would be proportional to the square of the frequency and the signal to noise ratio at the receiver would be a monatomically increasing function of frequency. We would then want to go to the highest possible frequency consistent with such things as our ability to point the sending antenna toward the receiver. All things being considered, it is most likely that optical frequencies are too high, while microwave frequencies are too low to be "best." It is a curious fact that on fundamental grounds one derives, as a most desirable region for communication, just that frequency range ; i.e., the far and middle infra-red, where no good transmitters and receivers presently exist.

References

1. C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, Urbana, Illinois, 1949.

2. J. P. Gordon, "Quantum Effects in Communications Systems," Proc. I.R.E., Vol. 50, pp. 1898-1908, Sept., 1962.

3. B. M. Oliver, "Comments on 'Noise in Photoelectric Mixing'," Proc. I.R.E. (Correspondence), Vol. 50, pp. 1545-1546, June, 1962.

Lasers: Principles, Processes and Potentialities (Abstract)

by C. Alley

The recent achievement of controlled amplification by stimulated emission in the range of optical frequencies was placed in the context of the development of our insights into the matter of electromagnetic radiation and its interaction with matter and relation to other developments in resonance physics and quantum electronics by a brief historical review. A short summary of some of our present concepts of this interaction of material interims of the quantum theory of the electromagnetic field, including the division and the uses of creation and annihilation operators for photons, was presented.

The various methods used so far to achieve amplifying conditions was discussed and the optical resonance configuration used to provide mode selection and feed back for substained oscillation was described. Some of the significance through numerical parameter characterizing the performance of different types of lasers was also presented.

A number of actual and potential applications of the lasers was mentioned along with a few of the problems associated with communications of long distances.

A cw helium-neon laser was set up to demonstrate some of the properties of its radiation, including a simple homodyne modulation and demodulation experiment.

Middleton (David), Concord, Mass.

15U4402

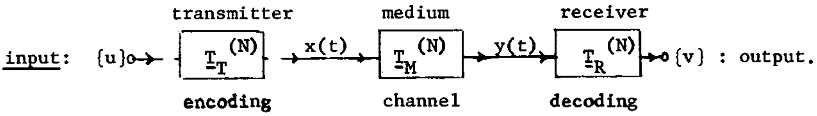
t. SPECIALIZED TOPICS III,

Decision Processes in Communication Theory

by David Middleton In ... p 74-76 refs

Communication theory, from the modern viewpoint, is a discipline, and an art, based on statistical concepts. The principal aims of such a theory are (1) to determine optimum systems; (2) their expected performance, and (3) to compare their structure and performance with suboptimum systems for the same purpose. Probability methods and statistical techniques provide the principal analytic tools for accomplishing these aims. A particularly powerful approach is based on the methods of statistical decision theory[1], which relates the values, or costs of decisions, to the decision themselves and the probabilities of their occurrence in the face of uncertainty. The following, thus, is a brief outline of the major elements of the present discussion.

Communication is a statistical process. We illustrate this with the basic single-link communication link shown in the figure below, which schematically relates a set of input messages



(or decisions), {u}, to a corresponding set of output messages (or decisions), {v}. Thus we have

$$\{v\} = \underline{T}_R^{(N)}\ \underline{T}_M^{(N)}\ \underline{T}_T^{(N)}\{u\},$$

where $\underline{T}_T^{(N)}$, $\underline{T}_M^{(N)}$, $\underline{T}_R^{(N)}$ are the operation, or transformations, respectively, introduced by the transmission process, the mediums, and the receiver. The superscript (N) denotes the insertion of noise in some fashion into the communication link at that particular stage. In practice, we have some powers to adjust $\underline{T}_T^{(N)}$, $\underline{T}_R^{(N)}$, but little or none to modify $\underline{T}_M^{(N)}$, which puts a fundament limit on the accuracy of the communication process, since $\underline{T}_M^{(N)}$ and $\underline{T}_R^{(N)}$ always inject noise into the system.

The expected, or average performance of the reception process, $\underline{T}_R^{(N)}$, (which we shall consider primarily in what follows) is measured by the <u>average</u> <u>risk</u>.

$$R(\sigma,\delta) = \int_{\Omega} \underline{dS}\ \sigma(\underline{S}) \int_{\Gamma} F_n\ (\underline{V}|\underline{S}) \int_{\Delta} C(\underline{S},\underline{\gamma})\ \delta(\underline{\gamma}|\underline{V})\ \underline{d\gamma},$$

where

$\underline{\gamma} = (\gamma_{11} \cdots \gamma_n)$ = set of possible decisions;

$\underline{V} = (V_{11} \ldots V_n)$ = received data vector ; $V_i = V(t_i)$, $i = 1,\ldots,n$, on an interval $(0,T)$;

$\delta(\underline{\gamma}|V)$ = probability, or probability density, of the decisions $\underline{\gamma}$, given the data $\underline{V}$; this is the <u>decision</u> <u>rule</u>.

$C(\underline{S},\underline{\gamma})$ = a cost function, relating the signal $\underline{S}$ to the decision $\underline{\gamma}$;

$\underline{S} = (S_{11} \ldots S_n)$, signal vector; $S_i = S(t_i)$, etc.

$\sigma(\underline{S})$ + a priori d.d. of $\underline{S}$

$F_n(\underline{V}|\underline{S})$ = the conditional d.d. of $\underline{V}$, given $\underline{S}$.

We distinguish two principal classes of decisions, $\underline{\gamma}$:

(a) <u>signal</u> <u>detection</u>: here $\underline{\gamma} = \gamma_1$: "yes" - a signal is present, or $\gamma = \gamma_o$: "no" - only noise occurs, and $\underline{\gamma} = (\gamma_o, \gamma_1)$.

(b) <u>signal</u> <u>extraction</u>: this is a measurement process, where "how much" is asked: we wish the magnitude of $\underline{S}$, or its waveform, etc.

<u>Detection</u> is recognized as the communication equivalent of <u>hypothesis</u> <u>testing</u> in statistical theory, while <u>extraction</u> is the counterpart of <u>statistical</u> <u>estimation</u>.

System structure is embodied in the decision rule δ. For optimum systems we seek a $\delta \rightarrow \delta^*$ which minimizes the average risk $R(\sigma,\delta)$, e.g.

min R = R*. A variety of maxima are possible. In detection, δ is a probability, and $C(S,\gamma)$ can be represented as a matrix of preassigned (constant) costs. For extraction, on the other hand, δ is a probability density, and the cost function becomes more suitable function of the decision δ and the quantity being "extracted", or estimated.

These above notions can be thus illustrated by simple examples involving detection and extraction; as they occur in radar and communication situations. (Details of the models and the techniques are given in Chapters 19, 20, 21 of Ref. 1. Some generalizations of these ideas are discussed in Chapter 23, Ref. 1.)

The treatment here concludes with a review of some of the advantages and difficulties of this general approach. The role of a priori information is examined, along with various ways of handling its presence and absence; the choice of cost functions is considered and the choice of the criteria of optimality. It is emphasized that these involve subjective notions by which it is thus possible to relate the "value" of a process and its outcome to the interactions of the physical world in which the process occurs. The decision process here is specifically incorporated into the data handling operations and, in fact, noticeably affects these operations. As indicated above, the decision theory approach is useful in radar, communications, underwater sound, etc. Important areas of future application are: i)adaptive systems; ii) pattern recognition; iii) multiple decision and sequential operation, etc.

Bibliography

1. Middleton, D. "An Introduction to Statistical Decision Theory," McGraw-Hill (1960); Chapter 18.

Optimum Modulation Methods for Data Encoding

(Abstract)

by R. W. Rochelle

A number of encoding techniques are available which allow the improvement of signal-to-noise ratio by an increase in the channel bandwidth. One of these, pulse-frequency modulation (PFM), has found use as an information encoding technique in satellites and space probes which are limited in both power and weight.

Examples were given of the design of a typical PFM telemetry system as a function of the format, information rate, accuracy, precision, and state-of-the-art components. Results from previous satellite flights were shown to illustrate the noise immunity performance that can be obtained.

Dr. Robert W. Rochelle is head of the Flight Data Systems Branch, Spacecraft Technology Division, the Goddard Space Flight Center, Greenbelt, Maryland. He holds a M. E. degree from the Yale University, New Haven, Connecticut and received the Ph. D. degree at the University of Maryland, College Park, Maryland in June 1963.

0852392 [Electronics Communications, Inc., Timonium, Md.

SPECIALIZED TOPICS IV,

Millimeter Waves

Donald D. King In ... p 78-92 refs

Introduction

The transition from microwaves to millimeter waves has been pedestrian - both slow and unimaginative when compared to the more spectacular progress in the optical region. In effect, the old infrared gap between radio and light has been bridged, and the millimeter region is now left behind. However, there remains much that is of interest in this spectral region for both the scientist and the engineer.

From a scientific point of view, there are several clear cut areas of interest. Most obvious, perhaps, is the fact that a vast number of spectral lines lie in the millimeter bands. Active and passive mapping of celestial bodies is another area of millimeter wave applications. High resolution coupled with different emission and reflection properties set millimeter mapping apart from both optical and microwave techniques. High energy plasma diagnostics, the observation of the sun and radio stars, and the analysis in depth of planetary atmospheres, are other scientific uses of millimeter waves.

From an engineering point of view, the usefulness of millimeter waves for communications and tracking depends on three distinct factors - 1) dimensions of the wavelength, 2) quantum noise, 3) atmospheric propagation. From the point of view of dimensions, it might appear that optical wavelengths are best for long range communication and tracking. This is not necessarily the case since the very small beam angles required for optical wavelengths introduce severe acquisition and tracking problems.

A more fundamental limitation in the optical region is quantum noise. The total noise power per unit frequency is

$$W(f) = \frac{hf}{\exp(hf/kT) - 1} + hf$$

In the microwave and millimeter region, it is generally true that

$$kT << hf$$

Then $W(f) = kT$

In contrast, in the optical region, the noise is proportional to the frequency. $W(f) = hf$ Depending on the temperature, the transition to frequency dependent noise occurs in the millimeter band.

A convenient formulation is

$$\frac{hf}{kT} = \frac{1}{20}\frac{f(Gc)}{T(^{o}K)}$$

Atmospheric propagation factors are summarized in Figures 1 and 2. Important features are the well-defined water-vapor and oxygen lines, their sharpening with altitude, the effect of humidity, and the relatively low vertical attenuation in the windows. At 4 km, the lower curve in Figure 1, the windows are significantly larger. Data above 150 Gc is very uncertain.

The state of the art in millimeter technology may be divided into generator, detector, and waveguide technology; these are now considered in turn.

Generators

There are three general classes of millimeter wave generators -

1) conventional electron tubes, such as klystrons, magnetrons, and backward wave oscillators
2) harmonic generators and related devices
3) megavolt electron beam schemes

Of these, conventional tubes are far and away the most important from a practical point of view.

The most successful modification of resonant designs has been the O-type carcinotron or backward wave oscillator. A major advantage of this type of tube is the separation of the three major parts - gun, interaction structure, and collector. The O-type carcinotron is a linear beam device, which might be considered as the outgrowth of stacking a number of coupled klystron cavities along a single electron beam.

A summary of the presently available electron tube generators is shown in Figure 3. This represents the best available output. However the relatively short life, high cost, and elaborate power supply restrict many of the best tubes to laboratory use. On the basis of past performance, it is reasonable to expect great improvement in these subsidiary factors. In terms of ultimate limits of the present designs, perhaps 600-800 Gc is a safe estimate for useful outputs.

The characteristics of the generators described improve rapidly at frequencies in the lower portion of the millimeter region. The expedient of using a convenient and relatively powerful source at lower frequency followed by a harmonic generator is therefore attractive. The use of frequency multipliers at lower frequencies has been highly successful. Here the term multiplier is used advisedly; only power at the desired harmonic is generated, the power in other harmonics being reflected by reactive terminations or idlers. The varactor or variable reactance element itself also does not absorb appreciable power.

As might be expected, such favorable characteristics have not been achieved at millimeter wavelengths. This is partly because of higher circuit losses, but the major cause is the poorer figure of merit of the varactor itself. This figure of merit is the cut-off frequency.

$$f_c = 1/2 \pi C_m R_s$$

The highest value of f_c so far reported is 800 Gc; commercially available units are generally below 150 Gc in cut-off frequency. Consequently varactors, which are PN junction devices, have not yet been used extensively on the millimeter region.

The basic problem lies in making a sufficiently small, stable PN junction. For this, finely pointed wires are welded or cemented to the active material, usually Gallium Arsenide - N-type. As the wavelength approaches 1 mm or less, the resistive terms in the equivalent circuit become dominant, and the junction behaves more like a variable resistance. The upper limit of efficiency for resistance variation has been derived by Page.

$$\frac{P_n}{P} \leq n^{-2}$$

Efficient coupling to the junction at both input and output frequencies is the most important property of a harmonic generator. More flexibility in coupling, and the best performance have been achieved with crossed guide arrangements. In general the second harmonic is 15-20 db below the fundamental. Higher harmonics are down in decreasing amounts; the difference between say 5th and 6th being only 3-4 db.

Related to the diode harmonic generator is the tunnel diode oscillator. Burrus has obtained measurable outputs to 3 mm wavelengths from tunnel diode oscillators. Here the limitation again appears to be junction size. Therefore, there is little hope for appreciable power output from these devices. However, low power suitable for local oscillator use may be attainable.

Other quantum mechanical methods for obtaining millimeter wave energy are being explored. The most obvious, the three level maser, is hampered by the need for pumping at a higher frequency.

Another mechanism related to masers is the use of a two level quantum system. Pumping at a submultiple of the transition frequency will produce output near the transition frequency, provided sufficient pump power is applied to make the off-diagonal or perturbations terms in the Hamiltonian significant. This mechanism is appropriately called harmonic generation by non-linear quantum susceptibility.

Non-linear effects also exist in ferrites, and have been used to generate harmonics in the millimeter band. A major advantage of ferfite harmonic generators is their ability to handle high peak power input. Outputs up to 50 watts peak at 2 mm have been obtained by ferrite multipliers.

The non-linear elements considered so far are in the low and medium power range. In the high power range, megavolt electron beams represent a highly non-linear element for producing harmonics in the millimeter region. In principle, very tightly bunched beams can be obtained at relativistic velocities. Such beams can then be used to excite high harmonics in cavities, by Doppler effect, or by Cerenkov effect.

In summary, harmonic generators using diodes are a prime source for low-level signals. Useful power for some laboratory measurements and for local oscillators are available at wavelengths down about .5 mm. Ferrite harmonic generators have given high peak power outputs where adequate fundamental power was available. Quantum mechanical oscillators and relativistic electronics are of little practical use at present. In terms of the past rate of progress, the last classification seems particularly unpromising.

Detectors

It is again convenient to distinguish three classes of millimeter devices useful for detection. As in the case of generators, relatively conventional microwave techniques have given some of the best results for detectors in the millimeter region. Several specific advances have resulted in order-of-magnitude improvements in sensitivity and maximum frequency for superheterodyne receivers in the millimeter region.

1) New semiconductors for point-contact mixers

2) Harmonic mixing

3) Low noise, broadband i.f. amplifiers

The characteristics of the three types of diodes are shown in Figure 4. The points to note are the curvature at the origin, which is greatest for the backward diode, and the negative resistance region of the tunnel diode.

The negative resistance region of the tunnel diode is, of course, the most spectacular new property. At lower frequencies it permits mixing with conversion gain. In the millimeter region, operation in the negative resistance region has not been achieved so far. The highest frequency at which amplification has been observed with tunnel diodes is 85.5 Gc. Stable mixer operation in the negative resistance region is considerably more difficult because of the multiple terminals involved. However, the possibility of at least reducing the conversion loss with tunnel diode mixers is one of the favorable prospects.

The backward diode is a limiting low-current version of the tunnel diode, in which the negative slope has disappeared. In return, the sharp curvature at the origin and low impedance level have given excellent sensitivity and superior noise characteristics.

The optimum local oscillator power for all these small junction area devices is low - of the order of 1 mw or less, and is lowest for tunnel and backward diodes. This has permitted the use of harmonic mixing, in which the local oscillator signal is injected at a submultiple of the desired frequency, and harmonics are generated within the mixer. Performance comparable to that of direct mixers is obtained.

In heterodyne detection, local oscillator noise can significantly raise the noise figure. In microwave receivers, local oscillator noise is generally cancelled in a balanced mixer. At millimeter wavelengths, it is difficult enough to realize an efficient crystal and waveguide configuration, without the added requirement of producing a dual balanced device. Fortunately, another expedient remains which has proven very effective.

The noise spectrum of local oscillators is centered about the oscillation frequency, and may extend over many megacycles. However, at 100 or more megacycles from the oscillator frequency, the residual noise is negligible. The use of a microwave intermediate frequency is therefore desirable in millimeter receivers. The availability of suitable low noise broadband amplifiers in the UHF and microwave region has contributed a great deal to millimeter receiver design.

Alternative to heterodyne receivers, are direct detectors which include crystal detectors followed by wide band video amplifiers. In addition, various infrared detectors such as Golay cells, bolometers, and cooled semi-conductors are available. Of these, the photoconductive detector described by Putley, and the semi-conductor bolometer of Low, the Putley detector has a detectivity

$$D^* = A^{1/2}/P_N = 2 \times 10^{11} \text{ cm cps}^{1/2}/\text{watt}$$

or a noise equivalent power

$$P_N = 10^{-11} \text{ watts/cps}^{1/2}$$

It consists of relatively pure In Sb biased with about 5 k gauss magnetic field at 77°K. Very shallow impurity levels are evidently required to produce photoconduction with photons of wavelengths up to 8 mm.

The Low bolometer uses cooled gallium doped germanium. When operated at 2.15°K its noise equivalent power is

$$P_N = 5 \times 10^{-13} \text{ watts/cps}^{1/2}$$

The acceptance band of these detectors is not accurately known, but is presumably quite wide - of the order of 10-100 Gc or 1/3 f. This is usually best for radiometric purposes, but not desirable for coherent signals. A direct comparison of heterodyne or linear receivers with square law devices is impossible in general. This is because the input frequency bandwidth enters as a square root for linear receivers, and directly in square law devices. Certainly, for narrow band applications, the heterodyne receiver is at an advantage. For radiometric service, the best that can be done is to widen the intermediate frequency bandwidth as much as possible and accept both upper and lower sidebands.

Masers have just begun to appear in the millimeter region. Traveling wave designs with appreciable bandwidth hold the greatest promise for millimeter detectors of the future. Rutile (chromium doped titania) has been used successfully in a traveling wave 8 mm maser pumped at 4 mm.

Comparative sensitivity calculations for three types of millimeter radiometers are given in terms of a minimum detectable temperature difference ΔT_m, and are as follows:

Harmonic mixing superhet at 2 mm

$$\Delta T_m = F T_o / \sqrt{B_{if} \times \tau}$$

$$= \underset{(25\ \text{db})}{300} \times 300 / \sqrt{3 \times 10^9 \times 1} \quad = \quad 1.6^o$$

Rutile maser at 8 mm

$$\Delta T_m = \underset{(1.6\ \text{db})}{1.45} \times 300 / \sqrt{75 \times 10^6 \times 1} = \quad .05^o$$

Cryogenic bolometer at 300 Gc

$$\Delta T_m = P_N / k\, B \sqrt{\tau}$$

$$= 10^{-12} / 1.38 \times 10^{-23} \times 100 \times 10^9 \times 1 \quad = \quad 0.7^o$$

where

F = radiometer noise figure

T = reference temperature

B_{if} = I.F. bandwidth

τ = output time constant

B = input bandwidth

In summary, for narrow band service, maser and superheterodyne receivers are most sensitive. For radiometers, the wide acceptance bands of direct detection receivers may provide equal or better sensitivity, especially at the shortest wavelengths.

Waveguides

To achieve low loss at millimeter wavelengths, the energy in waveguides must be spread over a larger cross-section. Accomplishing this without harmful effects from unwanted modes is difficult. The principal methods for doing so are:

1) overmoded hollow metal guide

2) trough guide

3) surface waveguides

4) beam guides

The attenuation properties of these guides are displayed in Figure 5. The lower loss guides generally require a larger cross-section. The lowest losses occur for very loosely bound waves, either interior, TE_{01}^{o}, or surface, HE_{11}. In both cases, bends are difficult to achieve; conversion to unwanted modes occurs in the hollow tube, and radiation takes place from the surface waveguide. Close tolerances to preserve uniformity of cross-section are also required to prevent these effects on straight runs.

Conversion to unwanted modes is more serious than radiation loss, in the sense that it may introduce reflections and delay distortion. On the other hand, the highly dispersive nature of the dielectric rod HE_{11} wave is also objectionable. Instead of using a single mode, which is maintained either as a dominant mode, or by mode suppression, it is also possible to accept multimode operation under certain conditions.

In conclusion, it appears that the present state of generators, detectors, and waveguides leaves much for future development. The prospects for improved detectors and waveguides are good. On the other hand, useful coherent sources over the .1 - 1.0 mm band are relatively remote. The existence of clearcut and urgent applications in various frequency regions would no doubt give impetus to the development of suitable generators.

REFERENCES

1. "Atmospheric Absorption of 10-400 kMcps Radiation," E. S. Rosenblum, Microwave Journal, pp. 91-96; March 1961.

2. "Millimeter Wave Generation," Coleman and Becker, IRE Trans., Vol. MTT-7, pp. 42-61; January 1959.

3. "Millimeter Wave Tubes," W. W. Teich, Electronics, pp. 37-44; May 25, 1962.

4. "The Nature of Millimetric Tubes," M. de Thomasson, Microwaves, pp. 39-42; October 1962.

5. "Nonlinear Effects in Ferrites," C. H. Townes, Ed., Columbia, pp. 314-323; 1960.

6. "Millimeter Wave Harmonic Generation in Ferrites," C. S. Gaskell, Proc. IRE 50, p. 326; March 1962.

7. "Microwave Radiometry," D. B. Harris, Microwave Journal, pp. 41-46, 47-54; April, May 1960.

8. "Submillimeter-Wave Radiometry," M. W. Long, W. K. Rivers, Proc. IRE 49, pp. 1024-1027; June 1961.

9. "High Sensitivity 100-300 Gc Radiometers," M. Cohn, F. L. Wentworth, J. C. Wiltse, to be published in Proc. IEEE.

10. "Infrared Detectors: A Review of Operational Detectors," R. F. Potter, W. L. Eisenman, Applied Optics, I, 5, pp. 567-574; September 1962.

11. "Quasi-Optical Surface Waveguide for the 100-300 Gc Region," F. Sobel, F. L. Wentworth, J. C. Wiltse, IRE Trans., MTT-9, pp. 512-518; November 1961.

12. "Some New Aspects of Laser Communications," G. K. Megla, Applied Optics, Vol. II, pp. 311-315; March 1963.

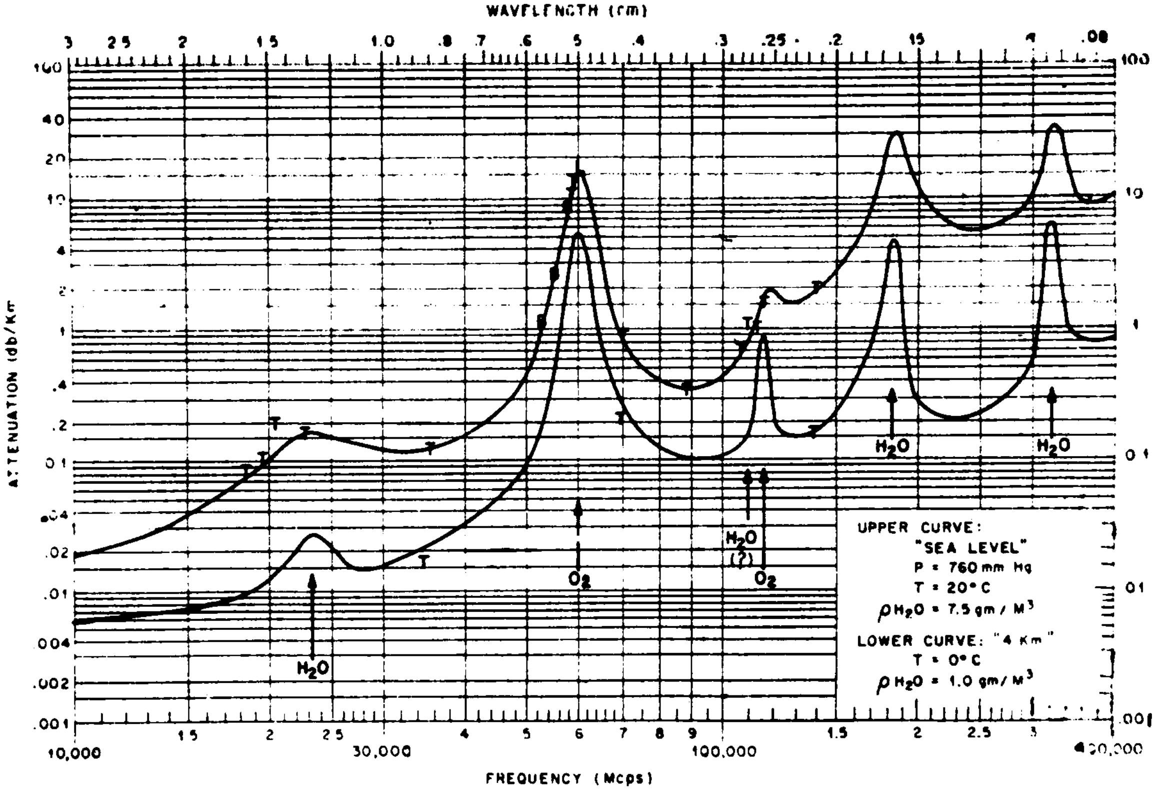


Figure 1
Horizontal Propagation
Characteristics after
Rosenblum
(Microwave Journal)

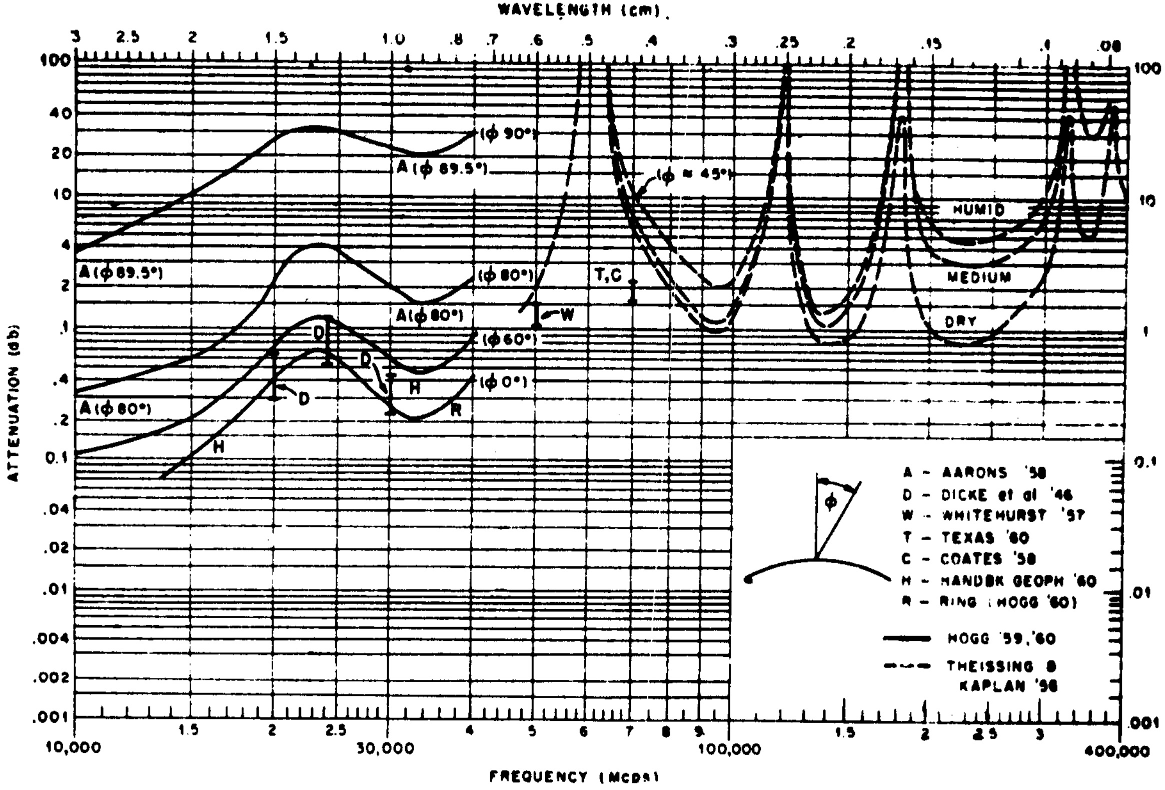
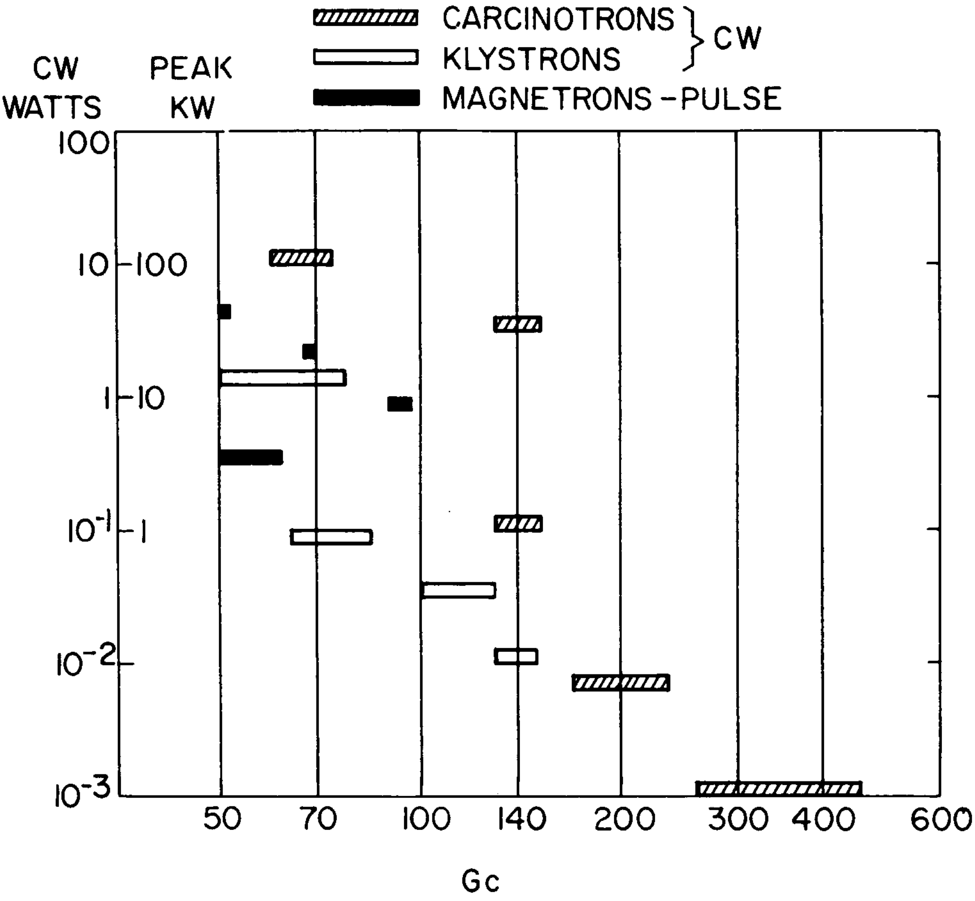
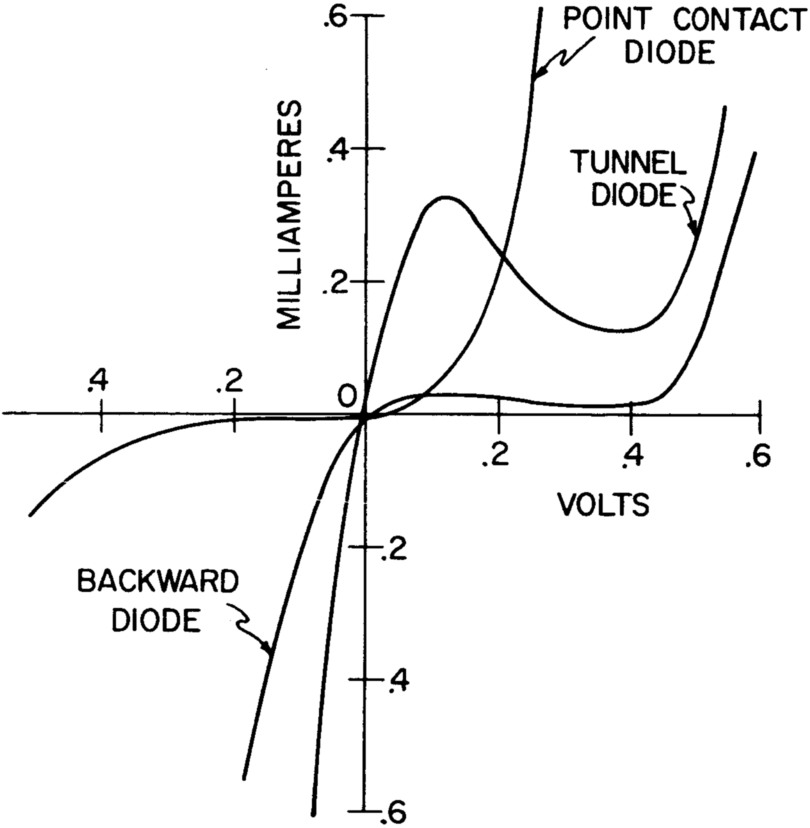


Figure 2
Vertical Propagation
Characteristics after
Rosenblum
(Microwave Journal)



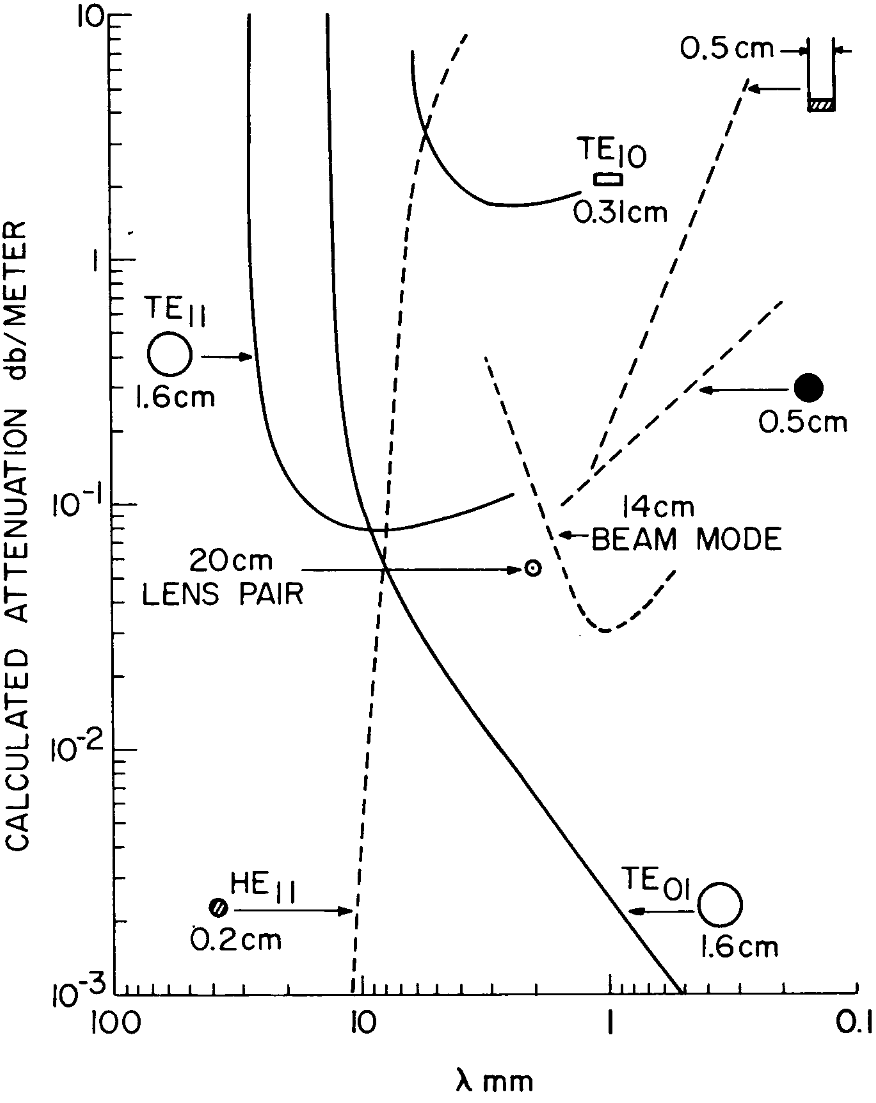
MILLIMETER WAVE OSCILLATORS

Figure 3



DIODE CHARACTERISTICS

Figure 4



LOSS IN MILLIMETER WAVEGUIDES

Figure 5

Princeton U., N.J. Plasma Physics Lab.

A885953

t, Plasma Waves (CMA Diagram)

by William M. Hooke In... p93-101 refs (see...)

One of the major problems in the study of plasma waves has been the profusion of names given to relatively few modes. This ambiguity has led to considerable confusion even among workers in the field of plasma waves.

The CMA (Clemmow-Mullaly-Allis) diagram[1,2,3,4] is used to classify and describe the various types of waves. The theoretical model on which the diagram is based is very simple: no collisions, low temperature, no boundaries. There are, however, no limits on the wave frequency, ω, the unperturbed magnetic field, B_o, or the density, N_e. A plane monochromatic wave of wave number $\vec{k}$ is assumed to propagate at an arbitrary angle, θ, with respect to $\vec{B}_o$.

The dispersion relation for this model was first obtained by Åström:[5]

$$\tan^2\theta = \frac{-P(n^2-R)\ (n^2-L)}{(Sn^2-RL)\ (n^2-P)} \qquad (1)$$

where for a plasma containing electrons of charge e and ions of charge Ze $\quad P \equiv 1-\alpha$

$$\alpha \equiv \frac{\pi_e^2 + \pi_i^2}{\omega^2} \qquad \pi_e^2 \equiv \frac{4\pi N_e e^2}{M_e}$$

$$\pi_i^2 \equiv \frac{4\pi N_i Z_i^2 e^2}{M_i}$$

$$R \equiv 1 - \frac{\alpha}{(1 + \frac{\Omega_i}{\omega})(1 - \frac{\Omega_e}{\omega})} \qquad \Omega_i = \frac{Z_e B_o}{M_i c}$$

$$\Omega_e = \frac{eB_o}{M_e c}$$

$$L \equiv 1 - \frac{\alpha}{(1 - \frac{\Omega_i}{\omega})\ (1 + \frac{\Omega_e}{\omega})}$$

$$S \equiv 1/2\ (R + L)$$

$$N^2 = \frac{k^2 c^2}{\omega^2}$$

The meaning of R and L becomes clear when θ is set equal to zero. The solutions of equation 1 are then $N^2 = R$ and $N^2 = L$. Thus R and L are just the square of the index of refraction for waves propagating parallel to the magnetic field.

By inspection of the equations, we see that n^2 as a function of Θ is determined if we specify the quantities $\frac{\Omega_e}{\omega}$ and $\frac{\pi_i^2 + \pi_e^2}{\omega^2}$. This fact is the basis for the CMA diagram.

The diagram is a two-dimensional plot with $\frac{\Omega_e^2}{\omega^2}$ and $\frac{\pi_e^2 + \pi_i^2}{\omega^2}$ being the ordinate and abscissa.

The diagram is divided into 13 regions by seven lines. Each of the lines represents a cut-off ($k^2 = o$) or a resonance ($k^2 \to \infty$). These lines separate regions containing different mode types.

The approximate equations (neglecting terms of order $\frac{ZM_e}{M_i}$) for these lines are:

1. Plasma cut-off $\omega \approx \pi_e$
2. Ion cyclotron resonance $\omega = \Omega_i$
3. Electron cyclotron resonance $\omega \,\square\, \Omega_e$
4. Lower hybrid resonance $\omega^2 \approx \dfrac{\Omega_i \Omega_e}{1 + \dfrac{\Omega_i \Omega_e}{\pi_i^2 + \pi_e^2}}$
5. Upper hybrid resonance $\omega^2 \approx \Omega_e^2 + \pi_e^2$
6. Ion cyclotron cut-off $\omega \quad \dfrac{\Omega_i - \Omega_e \pm (\Omega_e^2 + 2\Omega_e\Omega_i + 4\pi_e^2)^{1/2}}{2}$
7. Electron cyclotron cut-off $\omega \approx \dfrac{\Omega_e + (\Omega_e^2 + 4\pi_e^2)^{1/2}}{2}$

In each of the 13 regions polar plots of phase velocity may be drawn. Since the physical problem is cylindrically symmetrical about B_o, these curves are surfaces of revolution about the axis defined by θ = o. These wave normal surfaces exhibit the following important characteristics:

1. At each point on the diagram there are at most two phase velocity surfaces.
2. Inside a region bounded by the seven lines, the phase velocity surfaces are bounded.
3. The phase velocity surfaces are symmetrical about the $\theta = \pi/2$ axis.
4. There are only three topologically different types of surfaces.
   a. Spheroid 0
   b. **Dumbbell** lemniscoid 8
   c. Wheel lemniscoid ∞
5. The topology of a phase velocity surface remains constant within a bounded region.
6. Two phase velocity surfaces constructed for a given point in parameter space (i.e., a particular value of $\frac{\Omega_e}{\omega}$ and α) will never cross. That is, one volume

> **is always contained within another, so that the labels "fast" and "slow" apply for a given mode at all angles.**

Thus, wave normal plots may be drawn at one point in parameter space, and these surfaces will qualitatively represent the type of surface that exists throughout that entire region of parameter space.

The upper right-hand corner of the CMA diagram (p. 38, Ref. 3, p. 30, Ref. 4) contains two surfaces, one "spheroid" and one "**dumbbell** lemniscate." When $\omega \ll \Omega_i$ these are the compressional and torsional hydromagnetic waves respectively. As $|\vec{B}_0|$ decreases $\omega \to \Omega_i$ and the lemniscoid represents what is sometimes called the ion cyclotron wave. At $\omega = \Omega_i$ this mode experiences a resonance. The spheroid passes this line unaffected, but at the lower hybrid resonance line V phase = 0 at 90°, and the spheroid has become a **dumbbell** lemniscoid. Note that this hybrid resonance occurs only at 90°. As B_0 decreases further the angle at which resonance occurs becomes less than 90°. This mode which will not propagate at angles around 90° is the whistler mode. As $\omega \to \Omega_e$ the "resonant angle" becomes smaller (we have now the "electron cyclotron wave") and finally, there is a resonance at 0°, and the mode disappears. The lower part of the CMA diagram (characterized by $\omega \gg \Omega_i$) contains the magnetoionic modes of Appleton and Hartree.

The left side of the CMA diagram ($\alpha < 1$) represents in general the region where the plasma is so tenuous that the displacement current term $\frac{\partial \vec{E}}{\partial t}$ is greater than the conduction current term $4\pi \vec{J}$. If this condition holds we would expect the wave to be much like an electromagnetic wave in free space. This is the reason that there are three regions on the $\alpha < 1$ side of the diagram that are topological spheres as would be expected for free-space electromagnetic waves. There are two regions on the diagram where these qualitative statements do not hold. These regions are near the cyclotron frequencies. Along the electron cyclotron cutoff line, for example, the conduction currents perpendicular to B_0 are large enough to cancel the displacement current (for the mode with the electric field polarized in the same sense as the electrons) and the resonance occurs.

Thus the CMA diagram is a method for classifying and describing plasma waves over a broad domain of parameter space. Though the model is very simple, the diagram represents an excellent framework on which to display the results of the more realistic theories.

References

1. Clemmow, P. C. and R. F. Mullaly: Dependence of the Refractive Index in Magneto-Ionic Theory on the Direction of the Wave Normal, "Physics of the Ionosphere: Report of Phys. Soc. Conf. Cavendish Lab.," p. 340 Physical Society, London, 1955.

2. Allis, W. P.: Waves in a Plasma, Sherwood Conference. Contr. Fusion, Gatlinburg, Apr. 27-28, p. 32, TID-7582 (1959). Also in Mass. Inst. Technol. Research Lab. Electronics Quart. Prog. Rept., 54: 5 (1959).

3. Allis, W. P., S. J. Buchsbaum, and A. Bers: "Waves in Anisotropic Plasmas," M.I.T. Press, Cambridge, Massachusetts, 1963.

4. Stix., T. H.: "The Theory of Plasma Waves," McGraw-Hill, New York, 1962.

5. Åström, E. O.: On Waves in an Ionized Gas, Arkiv. Fysik, 2: 443 (1950).

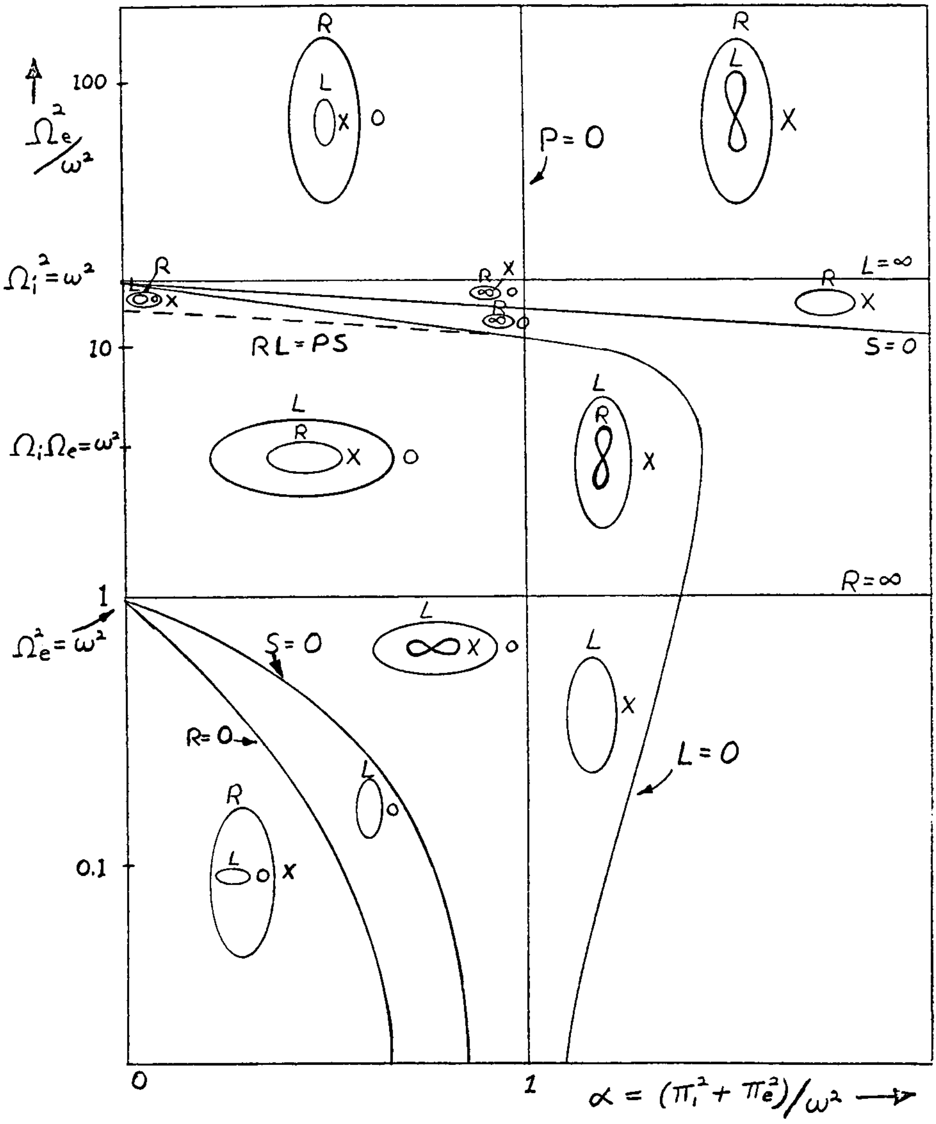


FIGURE 1

(Explanations on next page)

Figure 1: CMA diagram for a two-component plasma. (Reprinted with permission of author and publisher from figure 2-1, page 30 of The Theory of Plasma Waves, T. H. Stix, McGraw-Hill Book Company, Inc., New York, 1962). The ion-to-electron mass is chosen to be 4. Cross sections of phase velocity surfaces are sketched and labelled. For these sketches the direction of the magnetic field is vertical.

The labels R and L refer to the electric field polarization for propagation at $\theta = 0^{\circ}$. A wave is left handed polarized, (L), if the electric field rotates in the same sense as the gyration of positive particles in the steady magnetic field.

The labels 0 (ordinary) and X (extraordinary) refer to the form of the dispersion relation at $\theta = \pi/2$. The ordinary wave obeys $n^2 = P$ (independent of the magnetic field); the extraordinary wave obeys $n^2 = RL/S$.

Communication Systems
(Abstract)
by Albert Hedrich

The lecture dealt with computations of the performance data of active communication-satellite systems. Several satellite systems were considered and their performance compared. The effects of important parameters on the system characteristics were evaluated. These parameters included antenna and receiver noise temperatures, antenna gain, bandwidth, modulation techniques, satellite altitude and distribution, and ground- and satellite-transmitter powers. As examples, Relay, Telstar, and Syncom were considered.

Tracking Systems
(Abstract)
by Frederick Vonbun

An analysis was given of a new ranging system developed by the Goddard Space Flight Center for tracking satellites and space probes. Reasons for developing this tracking system in addition to the Minitrack System now in existence were presented. This Range and Range Rate Tracking System was described and analyzed in its simplest form. The errors associated with the tracking system were discussed in more detail. In addition, the discussion covered the results obtained in tracking experiments using the NASA SYNCOM Satellite which seemed to justify the development of this new tracking system.